A Similarity Measure for Indefinite Rankings

WILLIAM WEBBER
and
ALISTAIR MOFFAT
and
JUSTIN ZOBEL
The University of Melbourne, Australia

Note: Author's version; page numbers, editing vary from published version

Ranked lists are encountered in research and daily life, and it is often of interest to compare these lists, even when they are incomplete or have only some members in common. An example is document rankings returned for the same query by different search engines. A measure of the similarity between incomplete rankings should handle non-conjointness, weight high ranks more heavily than low, and be monotonic with increasing depth of evaluation; but no measure satisfying all these criteria currently exists. In this article, we propose a new measure having these qualities, namely rank-biased overlap (RBO). The RBO measure is based on a simple probabilistic user model. It provides monotonicity by calculating, at a given depth of evaluation, a base score that is non-decreasing with additional evaluation, and a maximum score that is non-increasing. An extrapolated score can be calculated between these bounds if a point estimate is required. RBO has a parameter which determines the strength of the weighting to top ranks. We extend RBO to handle tied ranks and rankings of different lengths. Finally, we give examples of the use of the measure in comparing the results produced by public search engines, and in assessing retrieval systems in the laboratory.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics—*correlation and regression analysis*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*experimental design*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation (efficiency and effectiveness)*

General Terms: Experimentation, Measurement, Human factors

Additional Key Words and Phrases: Rank correlation, probabilistic models, ranking

1. INTRODUCTION

Ranked, incomplete lists of items are encountered everywhere. Magazines list the most eligible bachelors; newspapers rank bestsellers; the registry reports the most popular boys' names for a year; and search engines rank documents by likelihood of relevance to a query. Such rankings share important characteristics. First, they are *incomplete*; that is, they do not cover all elements in the domain. The magazine lists the ten most eligible bachelors,

This work was supported by the Australian Research Council and by National ICT Australia (NICTA). NICTA is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.
Authors' addresses: Department of Computer Science and Software Engineering, The University of Melbourne, Victoria 3010, Australia. Email: {wew|alistair|jz}@csse.unimelb.edu.au.

not the entire population of marriageable men. Second, they are *top-weighted*; the top of the list is more important than the tail. The contention between best and second-best seller is more intense than between three-hundredth and three-hundredth and first. And third, they are *indefinite*; the decision to truncate the ranking at any particular depth is essentially arbitrary. The provider or user of the list could continue to enumerate items until the domain was exhausted, at least conceptually, but the cost involved is decreasingly justified by the utility obtained. A search engine might allow the user to scroll through two million results for the query "holiday ideas", but the user is unlikely to look beyond the first few dozen entries. These three characteristics of an *indefinite ranking* are related: because the ranking is top-weighted, value decays with depth; decaying value motivates a truncation of the list at some arbitrary rank; and truncation leaves the ranking incomplete.

Rankings are often compared. How closely do the bestseller lists of a newspaper and an online bookseller agree? Have tastes in boys' names changed much over the past ten years? And the goal of the comparison is frequently to infer the similarity of the processes which have generated the rankings. How alike are the results of two search engines over a series of queries? And how similar, therefore, are the collections and the retrieval algorithms of those engines? The objective and repeatable comparison of rankings requires a *rank similarity measure*. Such a measure needs to treat the peculiar features of indefinite rankings in a reasonable way. It must handle the appearance of items in one list but not the other. Differences at the top of the list ought to be given more weight than differences further down. The measure should not arbitrarily assign a cutoff depth, but be consistent for whatever depth is available from the list provider or reached by the user. And the measure should do all of the above while imposing a minimum of assumptions on the data, and none that violate the nature of indefinite rankings. A measure with these features qualifies as an *indefinite rank similarity measure*.

Given the ubiquity of indefinite rankings, it is surprising that there appear to be no indefinite rank similarity measures. There are many similarity measures on conjoint rankings (that is, where both lists consist of the same items). Tarsitano [2002] reviews thirty, and more have been proposed since then. Some metrics on conjoint rankings are top-weighted, and more can be made so. A few unweighted measures on non-conjoint rankings have been analysed, and a couple of top-weighted, non-conjoint measures have been described. But even amongst this last set, none of the existing measures properly handle the indefiniteness of indefinite rankings, instead assigning arbitrary cutoff depths and not maintaining monotonicity as these are varied.

In this article, we propose not merely a new, but (we argue) the first similarity measure that is appropriate for indefinite rankings, *rank-biased overlap* (RBO). This measure is based on (but is not tied to) a simple user model in which the user compares the overlap of the two rankings at incrementally increasing depths. The user has a certain level of patience, parameterized in the model, and after examining each depth has a fixed probability of stopping, modelled as a Bernoulli random variable. RBO is then calculated as the expected average overlap that the user observes in comparing the two lists. The measure takes a parameter that specifies the user's *persistence*, that is, the probability that the user, having examined the overlap at one rank, continues on to consider the overlap at the next. The product of these probabilities gives the probability that the user will reach a certain rank, defining the *weight* of the overlap to that rank. The weights are geometrically decreasing, but never reach zero, reflecting the indefinite nature of the ranking; moreover,

they are naturally convergent, so no normalization is required.

Under RBO, the overlap to each rank has a fixed weight. This provides an elegant mechanism for handling the incomplete rankings in a consistent way, without having to embed the depth of the evaluation in the metric. The (convergent) sum of the weights of the (potentially infinite) tail determines the gap or *residual* between the minimum and maximum similarity scores that could be attained on exhaustive evaluation. The minimum, maximum, and residual scores on partial RBO evaluation are all monotonic in depth. A point score can also be extrapolated.

Being based on simple set overlap, RBO handles non-conjointness in a natural way; indeed, it does not even assume that the two rankings are drawn from conjoint domains. Set similarity is a more natural basis for comparing indefinite and truncated lists than the more widely used one of correlation. In fact, RBO is a member of a family of weighted overlap measures, defined by taking the convergently weighted average of the overlap at different depths. Other weighting functions are possible within the same framework, including ones not based on a simple mathematical progression but derived directly from observed user or system behaviour.

There are many domains to which RBO could be usefully applied. In our experimental section, we concentrate on that of information retrieval. A common instance of indefinite rankings found in IR is the results lists returned, in decreasing order of estimated likelihood of relevance or utility, by retrieval systems. The lists of web pages returned by web search engines in response to user queries are the most familiar example. We give demonstrations of the uses of RBO in this environment, and of the problems that are encountered when instead applying measures that are not appropriate for indefinite rankings.

One reason for comparing the rankings of different retrieval systems is to explore how similar the two systems are, in the documents they index and the algorithms they use to determine which are relevant to a query. The comparison is *symmetric*; one system is not being measured against the other. In different circumstances, there may be an *objective* ranking (sometimes called the "gold standard") against which one or more *observed* rankings are being assessed. The objective ranking could, for instance, be returned by a precise-but-expensive retrieval algorithm, and the observed ranking by an algorithm that takes an efficiency-motivated short-cut. In this case, the researcher wishes to measure how far the observed ranking deviates from the objective. Frequently the assumption in an objective–observed comparison is that differences suggest a decrease in quality in the observed ranking, and the similarity measure is employed as a proxy for a full (and potentially expensive) retrieval effectiveness assessment. We give examples of both symmetric and objective–observed comparisons.

2. COMPARING RANKED LISTS

Internet users daily process ranked lists in the form of search engine results. A natural question to ask is how alike the rankings returned by different engines are. Figure 1 shows the results given by three popular web search engines to the query *'boarding school effect on children'*. How similar are these results? Are two of the lists closer to each other than to the third? A subjective judgment could be made for the results to a single query, but to generalize about the similarities of the engines themselves, many more queries would have to be considered. Some repeatable, easily computable, mutually comparable measure of result similarity is needed. What measure should be used?

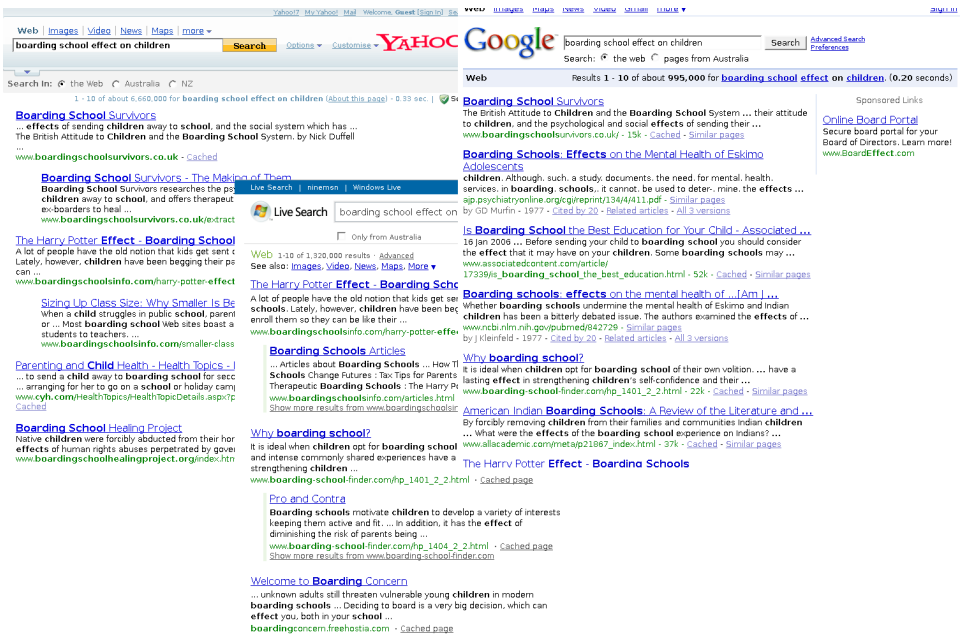


Fig. 1. Results received from the Australian portals of three public search engines to the query *'boarding school effect on children'*, issued on August 28th, 2008.

rnk	docid	sim
1	FBIS4-13392	6.44
2	FT931-12892	6.13
3	FT921-11935	5.66
4	FT933-7566	5.62
5	FT924-12615	5.49
6	FBIS4-59400	5.46
7	FT943-14288	5.31
8	FT941-373	5.30
9	FT923-12606	5.29
10	FBIS4-11824	5.29
…	…	…

(a) Full Evaluation

rnk	docid	sim
1	FBIS4-13392	6.44
2	FT931-12892	6.13
3	FT923-12606	5.29
4	FBIS4-11824	5.29
5	FBIS4-38863	5.24
6	FBIS4-46500	5.22
7	FBIS4-39925	5.19
8	FBIS4-46560	5.15
9	FBIS4-61085	5.00
10	FBIS3-55156	4.99
…	…	…

(b) 1000 Accumulators

rnk	docid	sim
1	FBIS4-13392	6.44
2	FT931-12892	6.13
3	FT921-11935	5.66
4	FT933-7566	5.62
5	FT943-14288	5.31
6	FT923-12606	5.29
7	FBIS4-11824	5.29
8	FBIS4-38863	5.24
9	FT942-2178	5.23
10	FBIS4-46500	5.22
…	…	…

(c) 400 Accumulators

Fig. 2. Runs returned by an experimental retrieval system to a test topic, under (a) full evaluation of index information; and (b, c) two different abbreviated evaluations. Each row is a document that the system has returned for the particular query. The second column gives the document identifier, by which the document is represented internally. The third column gives the similarity score calculated between each document and the query. The first column gives the document's rank in the result; the rank is determined by the similarity score.

Users encounter search engines on the public internet; researchers must wrestle with them in the lab. Figure 2 gives part of the output of a typical experiment. A shortcut to speed up query processing called query pruning is being examined. In query pruning, only the documents which, on an initial evaluation, seem most likely to be relevant are fully evaluated for relevance. Pruning speeds up processing, but at a possible cost in accuracy.

Two query pruning levels, one more severe than the other, are being tested against full evaluation. The document rankings produced by each method to depth 10 for a particular query are shown. Full evaluation serves here as an objective or "gold-standard" ranking, against which the two query pruning methods are compared. The researcher wants to quantify the impact on ranking fidelity that different levels of query pruning have, not just for this but for perhaps thousands of others of queries. How should the similarity of the pruned to the full-evaluation runs be measured?

Quantifying the similarities in each of the previous scenarios requires a measure of similarity between ranked lists of items. This might seem a well-understood problem in metrics, amenable to familiar rank correlation coefficients such as Kendall's τ [Kendall 1948]. There are, however, characteristics of these rankings that need to be carefully considered before choosing a measure to apply. Nor are these features in some way peculiar to search results; some or all of these features are observed in many other domains.

The first characteristic to be noted in the above lists is that the top of each ranking is more important than the bottom. The web pages returned by a search engine at the head of the ranking are more likely to be considered by users than those returned lower down. The documents ranked at the top of an experimental system's run will have the most impact on the retrieval effectiveness score the system achieves. More subtly, the gap between the estimated similarity of different documents to the query becomes narrower the deeper the ranking is examined; some of this effect can be seen in the similarity scores reported in Figure 2. A corollary of the *top-weightedness* of these rankings is that exchanges or perturbations in ordering at the top of the ranking are more significant than those at the bottom. It therefore follows that a desirable feature of a measure of similarity between top-weighted rankings is that it imposes a stronger penalty on differences at the top of the ranking than on differences further down.

The second characteristic of web page and document rankings is that they are *incomplete*, not providing a full ranking of their domains. As a result, such rankings are mutually *non-conjoint*, with some elements turning up in one ranking but not the other. Most rank similarity measures require the two rankings to be conjoint, and cannot be applied unmodified to non-conjoint rankings. Even amongst similarity metrics on incomplete rankings, the majority assume that the underlying full rankings exhaustively order a common domain, and hence are conjoint. For instance, a common approach to handling non-conjointness is to assume that items returned in one ranking by cut-off depth k, but not in a second ranking by that depth, are ranked at depth $k + 1$ in the latter ranking. But the assumption that the full rankings exhaustively order a common domain is not always valid. For example, a search engine may not have in its index at all a web page returned by another engine, due to differences in crawling policies and processes. In such cases, assuming that an unreturned item is placed at some unobserved rank is unsatisfactory; if the former search engine were aware of the web page in question, it might well rank it in the first position. In general, therefore, it is preferable for a metric on incomplete rankings to handle non-conjointness directly, rather than making assumptions about an underlying conjointness.

The characteristics of *top-weightedness* and *incompleteness* observed in these rankings are related to a third important characteristic, that of *indefiniteness*. The distinguishing features of an indefinite ranking are that only a prefix of the list is being considered; that the prefix covers only a small fraction of the list; and, most importantly, that the length of the prefix is essentially arbitrary. Longer or shorter prefixes could be considered. The choice of

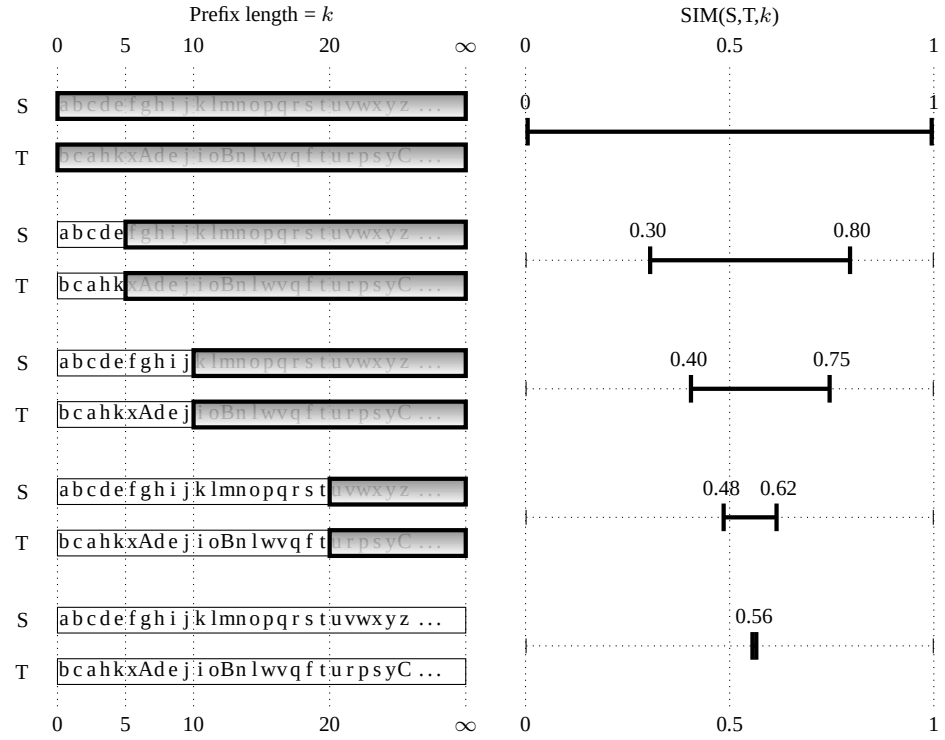


Fig. 3. Convergence of scores with more information. Before examining either of the rankings, their similarity score could range anywhere from 0 to 1. As the length of the examined prefix k increases, the range of the possible full similarity score decreases monotonically. These ranges bound the similarity score achievable on infinite evaluation.

prefix length could even depend upon the degree of similarity observed: greater fidelity of measurement might be required where rankings are similar, whereas less is needed where they are markedly different. And because multiple comparisons might be made in parallel, as when two search engines are compared on a number of different queries, and each comparison might have a different depth, scores should be comparable independent of depth. For these reasons, it is desirable that the measure chosen not have the depth of assessment embedded in it.

A ranking that has the qualities of top-weightedness, incompleteness, and indefiniteness described above, is referred to here as an *indefinite ranking*, and a measure of similarity between such rankings that meets all of the requirements outlined in the preceding paragraphs is referred to as a *similarity measure on indefinite rankings* or an *indefinite rank similarity measure*. Our aim in this paper is to show that existing rank similarity measures are not adequate *indefinite* rank similarity measures, and then to propose a new measure, rank-biased overlap, that is.

The central idea of our approach is to define a measure on the similarity of the full rankings, and then bound or estimate the full similarity value based on the list prefixes. After all, it is typically the similarity of the full rankings that is of interest, not just of their

	Unweighted	Weighted
Conjoint	Kendall's τ	Yilmaz's τ_{AP}
	Spearman's ρ	Iman-Conover's r_T
	Spearman's footrule	Shieh's τ_w
	Kolmogorov-Smirnov's D	Melucci's τ_*
	Carterette's d_{rank}	Blest's w
Non-conjoint	Fagin's τ_k, f_k, ρ_k, γ_k	Fagin's intersection metric
	Bar-Ilan's ρ, footrule	Bar-Ilan's M
		Buckley's AnchorMAP

Table I. Classification of rank similarity measures.

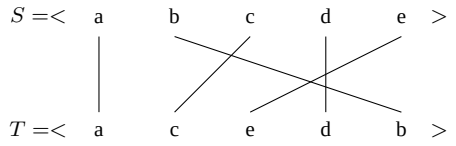
prefixes. The key to measuring full rankings from their prefixes is to choose a measure for which a partial, prefix evaluation bounds the value that a full evaluation would produce. The deeper the prefix that is examined, the narrower the bounds on the full score become. The idea is illustrated in Figure 3. With no elements examined, the similarity score between the rankings could take any value in the measure's legal range; say, anywhere between 0 and 1. After seeing the first 5 elements in each ranking, the possible scores of a full evaluation is narrowed to between, say, 0.3 and 0.8. And after extending the prefixes to depth 20, the range is narrowed to 0.48 and 0.62, as illustrated in the fourth segment of Figure 3. Technical details follow, but the principle is to choose a sequence of decreasing weights over the depths of the comparison, such that the sum of the weights is convergent; that is, so that the weight of the unseen, conceptually infinite tail of the lists is limited, and does not dominate the weight of the seen, finite prefix. Such a weighting scheme, besides being attractive mathematically, is justified representationally by the assumptions underlying indefinite rankings; that is, that the interest of the consumer of the ranking is sufficiently top-weighted for a truncated ranking to be satisfactory.

3. RANK SIMILARITY MEASURES

There are many rank similarity measures described in the literature. We categorize them according to the characteristics of indefinite rankings described in the previous section. Measures may be unweighted or top-weighted, and may require conjointness or support non-conjoint rankings. Non-conjoint rankings are sometimes referred to in the literature as top-k rankings; that is, prefix rankings evaluated to a fixed depth k. The existing measures are summarized in Table I and described in this section.

3.1 Conjoint, unweighted measures

The most widely used rank similarity measures are those that are unweighted and assume conjointness between the rankings. These predominantly fall into the class of *correlation* measures or coefficients. Correlation quantifies the type (positive or negative) and degree of relation between the two variates in bivariate, paired data. For instance, is height positively or negatively related to dancing ability, and how strongly so? If the observed variate pairs (in this example, people) are randomly sampled from a larger population, then the correlation in the sample can be used to infer the correlation in the population, and to test for the significance of the latter correlation. Since correlation can be either positive or negative, correlation coefficients typically range from -1 to 1, with -1 meaning perfect negative correlation (for rankings, in reverse order), 1 meaning perfect positive correlation (in identical order), and 0 meaning uncorrelated or "randomly" related [Gibbons and



$$
\begin{aligned}
\mathcal{C}_{ST} &= \{(a,b),(a,c),(a,d),(a,e),(c,d),(c,e)\} \\
\mathcal{D}_{ST} &= \{(c,b),(d,b),(e,b),(e,d)\} \\
C &= |\mathcal{C}_{ST}| = 6 \\
D &= |\mathcal{D}_{ST}| = 4 \\
M &= C - D = 2 \\
\tau &= M/P(n) = 2/(C+D) = 2/10 = 0.2
\end{aligned}
$$

Fig. 4. Example working of Kendall's τ.

Chakraborti 2003, Chapter 11].

One widely-used rank correlation coefficient is Kendall's τ [Kendall 1948]. To calculate τ, consider every pair of items from the set of items listed by the two (conjoint) rankings. Assume that there are no ties, that is, no two items have the same rank in either ranking (a variant of Kendall's τ handles ties, but will not be discussed here). Let C be the number of concordant pairs, where each ranking places the two items in the same order, and let D be the number of discordant pairs. Then $M = C - D$ is our basic statistic. The maximum value M can take for rankings of length n is the number of distinct pairs amongst n items, $P(n) = \binom{n}{2} = \frac{1}{2}n(n-1)$, and the minimum is $-P(n)$. Then, τ is derived as $M/P(n)$, and ranges from -1, indicating reverse order, to 1, indicating identical order. The value $P(n) - M$ can be understood as the number of adjacent pairwise swaps needed to arrange one ranking in the same order as the other, as in a bubble sort.

A working of Kendall's τ on two example rankings S and T is given in Figure 4. The set of concordant pairs is enumerated in $\mathcal{C}_{ST}$, while $\mathcal{D}_{ST}$ lists the discordant pairs. Discordant pairs can be found graphically by drawing a straight line between each item in S and the corresponding item in T, as is done in the figure; whenever two of these lines cross, the ordering of the respective items is discordant. The total number of pairs is the sum of C, the number of concordant pairs, and D, the number of discordant pairs, so τ is the proportion of these pairs that are concordant, linearly adjusted to the range $[-1, 1]$.

Kendall's τ has a direct probabilistic interpretation. Pick a pair of items ij at random from the set of $P(n)$ pairs. The probability p_c that ij are ranked in the same order in both rankings (that is, are concordant) is $C/P(n)$, and the probability p_d that ij are discordant is $D/P(n)$. A little algebra shows that $\tau = p_c - p_d$. Therefore, a τ of 0 indicates that a randomly chosen pair is as likely to be concordant as discordant. The rankings are then said to be uncorrelated. Furthermore, if the ranked items are assumed to be randomly sampled from a larger population of items, τ on the sample, which is sometimes denoted t, is an estimate of τ on the population. Inferential methods beyond point estimation from (sample) t to (population) τ are also possible, such as calculation of confidence intervals and testing of the null hypothesis of non-correlation, that is, that (population) $\tau = 0$.

Kendall's τ is widely used in the IR domain, and other fields, as a measure of rank correlation. Melucci [2007] and Yilmaz et al. [2008] list illustrative cases. But τ has none of the specific characteristics that we have set out for a measure of similarity on indefinite

rankings. First, it requires that the two rankings be conjoint. Second, it is unweighted, placing as much emphasis on disorder at the bottom of the ranking as it does on disorder at the top. Third, τ values are intrinsically linked to the depth of the ranking k. The contribution of the concordance of discordance of a single pair ij to the overall score is normalized by $P(k)$; as k increases, the significance of each disorder decreases.

Moreover, the concept of correlation itself is not helpful or meaningful when applied to indefinite rankings; that is, to rankings of which only the head is seen, and where the head is a small fraction of the entire (conceptual) list. Viewed through such prefixes, random and negatively correlated rankings will look essentially the same, having few or no common items in the observed head of the rankings. Indeed, for an indefinite rank similarity measure to be even considered, there must be an underlying presumption that the two rankings are strongly correlated, at least at the top of the ranking. What is being tested must be the departure, not from randomness, but from agreement. Thus, an indefinite rank similarity measure might be applied to the results of the same query on two search systems, to determine how different they are, since we would expect different engines to give similar results to the one query; but it would not be informative to apply it to the results of two distinct queries on the same search engine, to test how related the queries are, since we would expect the results of different queries to have little in common with each other in most cases. A basic implication is that an indefinite rank similarity measure should range not from -1, meaning negatively correlated, but from 0, meaning entirely dissimilar as far as can be seen, which is to say, disjoint. These objections to using the concept of correlation with indefinite rankings apply not only to Kendall's τ, but also to other rank correlation methods, and to the weighted and top-k measures derived from them, which is to say to the majority of the measures proposed in the literature.

Because indefinite rank similarity measures presume a strong relationship between the (full) rankings, testing for statistical significance becomes problematic. The standard null hypothesis in statistical tests on rankings is that the rankings (or rather the variates on the underlying population they are sampled from) are uncorrelated or randomly related, and significance is found for them having a (positive or negative) relation. The null hypothesis that the two rankings are, in contrast, identical is not helpful, as any finding of difference disproves this null hypothesis. It is true that confidence intervals on a rank correlation measure can be derived. Similarly, in the case where two rankings are being compared against a third, objective ranking, it is possible to test that one of the two rankings is significantly closer to the objective ranking than the other (see Cliff [1996, Chapter 3] for more details), although this method is not widely applied in the IR literature. But such confidence intervals tend to be very wide, meaning that observed sample correlation places weak bounds on inferred population correlation. In any case, the items in an indefinite ranking are not typically a random sample from a larger population; rather, the ranking itself is over the full population, but only the prefix is observed. Additionally, at least when considering the ordered lists of documents returned by retrieval systems, what is of interest is generally not the degree of relation on a given pair of rankings, that is, on the results to a particular query, but the similarity for all queries. Here, the (conceptual) random selection is not of documents, but of queries, and statistical inference can proceed along the standard lines for estimating population parameters from samples.

Other unweighted measures on conjoint rankings are available. The most widely used alternative is Spearman's ρ, which is the standard product-moment correlation, calculated

on ranks rather than magnitudes. Disorders in Spearman's ρ are penalized by the square of the distance of the disorder. In contrast, Spearman's footrule penalizes disorders by the unsquared distance. However, Spearman's footrule has problems in scaling, sensitivity, and analysis; see Kendall [1948] for more details. As noted above, the previous rank correlation measures are most naturally (though not only) suited to testing the null hypothesis of no correlation, whereas in many applications, the interest is rather in degree of departure from agreement. To address this, Melucci [2007] proposes the use of Kolmogorov-Smirnov's D, which is based on Kolmogorov's goodness-of-fit test. This test, though, only takes account of the single largest pairwise disordering; other disorderings in the list are ignored. All these unweighted measures on conjoint rankings share the inadequacies of Kendall's τ when applied to indefinite rankings. A more specialized rank similarity measure, d_{rank}, is proposed by Carterette [2009]. The measure requires that the ranked items are scored, and that these scores are aggregates of sub-scores over a common domain (such as systems ranked by scores over the one topic set). The measure takes account of score differences and correlations involved in discordant orderings. It is not top-weighted, and requires that the rankings be conjoint.

3.2 Weighted measures on conjoint rankings

It frequently happens with conjoint rankings that the top of the ranking is more significant than the bottom. A common way of comparing effectiveness metrics in IR, for instance, is by measuring the similarity between the rankings each metric induces over a set of retrieval systems. For these comparisons, the researcher will frequently care more about the ordering of good systems than that of bad. In such circumstances, a measure of similarity that gives greater weight to higher rankings may be desired. Such measures are often derived by adapting unweighted rank correlations.

Yilmaz et al. [2008] propose a top-weighted variant on Kendall's τ called τ_{AP}, based on the average precision retrieval effectiveness metric. The τ_{AP} measure has the following probabilistic interpretation. Denote the rankings to be compared as S and T. Randomly select an item i other than the top-ranked item in S. Next, randomly select another item j from those ranked higher than i in S. Then, see if i and j have the same order in ranking T. Let the probability of observing concordance in this random experiment be p_c, and the probability of observing discordance be p_d. Then $\tau_{AP} = p_c - p_d$. The similarity between τ_{AP} and τ is obvious, the only difference being in the method of selecting the items. Yilmaz et al. also demonstrate that if discordance is evenly spread throughout the rankings (not greater at the top than at the bottom), then $\tau_{AP} = \tau$. The top-weighting in τ_{AP} comes from the higher probability that item j is selected from the upper ranks of ranking S.

The τ_{AP} measure is not symmetric. The top-weightedness is defined solely on S, the ranking on which item j is selected above item i. Where S is the objective ranking, this lack of symmetry is acceptable, but if the rankings are of equivalent status, the lack of symmetry is problematic. Yilmaz et al. propose a symmetric alternative, which is the average of $\tau_{AP}(S,T)$ and $\tau_{AP}(T,S)$. To extend the probabilistic interpretation, the random experiment is modified to add the first step of randomly selecting a ranking to sample $j > i$ from. Unlike some other proposals, τ_{AP} has no parameter to set the degree or type of top-weightedness.

Regarding indefinite rankings, τ_{AP} satisfies the requirement of top-weightedness. It does not, however, handle incomplete, mutually disjoint rankings. Nor does it deal with

indefinite lists as laid out in Section 2: the depth of evaluation is implicitly embedded in the measure, and scores are not monotonic or bounded as the evaluation depth increases

Other weighted correlation measures on conjoint rankings are described in the literature. Iman and Conover [1987] apply Pearson's correlation coefficient not to raw ranks, as with Spearman's ρ, but to the Savage scores of the ranks. The Savage score S_i of rank i in a list of length n is $\sum_{j=i}^{n} 1/j = H(n) - H(i) \approx \ln(n/i)$, where $H(n)$ is the nth Harmonic number. Assuming no ties, their coefficient is calculated as:

$$r_T = \left(\sum_{i=1}^{n} S_{R_i} S_{Q_i} - n \right) / (n - S_1)$$

Because $S_i > S_{i+1}$, the coefficient r_T is top-weighted.

Shieh [1998] analyses τ_w, a class of weighted variants on Kendall's τ, where each pair of ranks can be assigned its own weighting. A suitable choice of weights makes this measure top-weighted. A subfamily of such measures, τ_*, is described and analysed by Melucci [2009]. The τ_* family itself generalizes τ_{AP} by allowing arbitrary weights to be assigned to the lower rank of rank pairs in the objective ranking; Melucci provides the probability distribution for the measures in the family. The τ_* family is non-symmetric, since one of the rankings is designated as the objective ranking. Blest [2000] defines a rank correlation w based on the difference in area between a polygon defined by the cumulative ranks of the observed ranking and a polygon defined by the cumulative ranks of the reversed objective order; this measure, too, is top-weighted. Quade and Salama [1992] survey earlier work on weighted rank correlations. None of these top-weighted measures directly handles incompleteness or indefiniteness.

3.3 Unweighted non-conjoint measures

The measures considered so far all assume that the two rankings are conjoint, that is, that every element occurring in one list also occurs in the other, and vice versa. They do not, unmodified, handle non-conjoint rankings. One way in which non-conjoint rankings occur is when longer, conjoint rankings are truncated to a fixed depth k. These truncated rankings are known as *top-k lists*.

Similarity measures on top-k and other non-conjoint rankings are frequently derived through the modification of a conjoint rank similarity measure. One such modification is simply to ignore non-conjoint elements. This approach is in general unsatisfactory, however, since it throws away information. For instance, if Kendall's τ were modified in this way, then the rankings $<$`ab???`$>$ and $<$`a???b`$>$, where `?` denotes a non-conjoint element, would be regarded as completely similar, when clearly they are not.

Rather than ignoring element i which appears in ranking S and not in ranking T, a more satisfactory approach is to treat i as appearing in ranking T at rank $k+1$ or beyond, where the depth of T is k. This agrees with the concept of top-k rankings, which assumes that the full domains are conjoint (that is, each element is ranked somewhere in the full list), but that only the top k positions are visible.

Placing unranked items below rank k is the approach taken by Fagin et al. [2003]. They adapt both Kendall's τ and Spearman's footrule in this way to handle top-k lists. For τ_k, the top-k version of τ, if element i appears in ranking S but not ranking T, it is assumed to be ranked beneath every item that does appear in ranking T. The only ambiguity occurs if elements i and j both appear in ranking S, but neither appear in ranking T. In this case,

Fagin et al. provide for a parameterizable penalty of between 0 (assumed concordant) and 1 (assumed discordant). They propose that the default value for this penalty should be 0, as this fixes the score for conjoint but reversed as close as possible to half way between identical and disjoint. A top-k version of Spearman's footrule, f_k, is similarly defined.

The measures τ_k and f_k are not top-weighted, but similar assumptions could be applied to top-weighted conjoint rank measures to derive weighted top-k measures. Weightedness makes the assumption of unlisted elements being ranked beyond rank k more complex, though. For instance, in τ_{AP}, when randomly selecting an item i and a higher-ranked item j, the question arises of whether the items beyond depth k are to be regarded as above or below each other. In particular, τ_{AP} does not (as currently defined) handle tied items, so the non-conjoint elements cannot simply be placed at rank $k + 1$. Instead, Yilmaz et al. [2008] propose that any such elements be excluded; but this loses information about implied misorderings, as described above.

The desideratum stated by Fagin et al. that conjoint but reversed top-k rankings should score roughly half way between identical and disjoint is not a compelling one. How close a relatedness reverse conjointness indicates depends on how large k is in relation to the full list size n. Moreover, conjoint but reversed to depth k is more a peculiarity than a meaningful characteristic for top-k lists, since by definition it cannot continue to be true if the evaluation is then extended to depth $k + 1$. Partly at fault is a desire to produce a measure that is similar in form to correlation measures on conjoint lists; having a negative score for a top-k measure is hardly meaningful. More fundamentally, though, correlation-based measures do not properly reflect the fact that these are indefinite rankings, and that the choice of k as the cutoff point is essentially an arbitrary one.

In addition to Kendall's τ and Spearman's footrule, Fagin et al. describe a top-k variant of Spearman's ρ. The treatment of non-conjoint elements is similar to that for the other methods; however, the resulting measure does not fall into the same equivalence class.

Goodman and Kruskal's γ is a correlation coefficient related to Kendall's τ, in which tied items are effectively ignored [Goodman and Kruskal 1954]. Fagin et al. also extend γ to the top-k case by regarding the pair ij both appearing in list S but neither appearing in list T as tied, and therefore ignoring it.

Bar-Ilan [2005] and Bar-Ilan et al. [2006] adapt Spearman's ρ and Spearman's footrule respectively to the top-k case by excluding non-conjoint elements (rather than treating them as occurring beyond depth k) and calculating the coefficients on the condensed lists. Bar-Ilan et al. point out the loss of information that condensing lists in this way entails.

3.4 Weighted non-conjoint measures

Most of the measures discussed so far have been founded upon correlation. When dealing with non-conjoint lists, it is also possible, and arguably more natural, to start instead from set intersection. A simple similarity measure on top-k lists would be the size of intersection or overlap between the two rankings, calculated as the proportion of the ranking length; that is, $|S \cap T|/k$. Of course, such a measure, while directly handling non-conjointness, takes no notice of ranking, and therefore is not top-weighted.

The idea of overlap can be extended by considering, not simply the overlap at depth k, but the cumulative overlap at increasing depths. For each $d \in \{1 \ldots k\}$, calculate the overlap at d, and then average those overlaps to derive the similarity measure. This measure is described by Fagin et al. [2003] and called the intersection metric, and was simultaneously discovered by Wu and Crestani [2003] and named average accuracy. We

d	$S_{:d}$	$T_{:d}$	$A_{S,T,d}$	$\mathrm{AO}(S,T,d)$
1	<a>	<z>	0.000	0.000
2	<ab>	<zc>	0.000	0.000
3	<abc>	<zca>	0.667	0.222
4	<abcd>	<zcav>	0.500	0.292
5	<abcde>	<zcavw>	0.400	0.313
6	<abcdef>	<zcavwx>	0.333	0.317
7	<abcdefg>	<zcavwxy>	0.286	0.312
n	<abcdefg...>	<zcavwxy...>	?	?

Fig. 5. Average overlap AO of two lists to increasing depths, along with their proportional overlap or agreement A at each depth. Average overlap continues to increase even as agreement decreases, and the value at depth k does not bound the value at arbitrary depth $n > k$. The notation used is described in more detail in Section 4.1.

refer to it as average overlap (AO). Because of its cumulative nature, AO is top-weighted: rank 1 is included in every subset, rank 2 in every subset but the first, and rank r in subsets r through k but not 1 through $r-1$. Thus, AO is the first of the measures we have examined that both handles non-conjoint lists and is top-weighted, and indeed is one of the very few described in the literature. Figure 5 gives a sample calculation.

Although average overlap is weighted and non-conjoint, and is closer to a satisfactory indefinite rank similarity measure than any of the previous alternatives, it fails our criteria for an indefinite measure because it is a measure not on infinite lists, but on their prefixes. One might be tempted to attempt the conversion of AO to an indefinite measure by conceiving of a score on the full lists, and then using the prefix evaluation to set bounds on it; that is, calculate AO@k (or a derivative) and use it to limit AO@∞. But such an attempt fails, due to the measure's non-convergence. The weight of the infinite tail always dominates that of the finite prefix, no matter how long the prefix is. A proof is given in Appendix A; intuitively, we see that each overlap to depth k has weight $1/k$ under AO@k, but weight $1/\infty$ under AO@∞. Thus, prefix evaluation sets no bounds on the full score: after comparing k elements, the AO@∞ score could still be anywhere in the range $[0, 1]$, not matter how large k is.

Average overlap has another peculiarity related to monotonicity in depth, which is that finding greater agreement with deeper evaluation does not necessarily lead to a higher score, nor finding decreased agreement to a lower one. For instance, in Figure 5, the elements newly revealed at depths 4 through 6 are all disjoint, yet the AO score keeps increasing. This counter-intuitive behaviour occurs because in calculating AO, the contribution of each overlap at depth d is only considered up to k, whereas in fact it continues to contribute up to n as n goes to infinity; increasing the evaluation depth k thus captures more of this residual contribution.

Bar-Ilan et al. [2006] describe and employ a measure M which is the normalized sum of the difference in the reciprocal of the ranks of each item in the two lists, with items not ranked in one list assumed to occur at depth $k + 1$ in that ranking. Like AO, this measure is top-weighted and handles non-conjointness, but is dependent on the cutoff depth k.

Buckley [2004] proposes the AnchorMAP measure, which is based upon the retrieval effectiveness evaluation metric, (mean) average precision (MAP). Retrieval evaluation metrics score a document ranking according to the relevance of the documents it contains. In AnchorMAP, one of the rankings under comparison is chosen as the objective ranking, and its first s documents are treated as relevant; Buckley suggests $s = 30$ as a reasonable value.

The MAP score of the other ranking is then calculated to depth k, based on these artificial relevance judgments. AnchorMAP is asymmetric. It is top-weighted, but weights are not fixed for ranks. The metric is non-monotonic both in s and k.

A referee of this work suggests an alternative mechanism, based on a rank-weighted evaluation metric such as discounted cumulative gain (DCG) [Järvelin and Kekäläinen 2002]. In a rank-weighted metric, each rank i is assigned a fixed weight w_i, and the document at that rank makes a contribution $w_i \cdot r_i$ to the ranking's effectiveness score, where r_i is the document's assessed degree of relevance. A similarity measure between two rankings S and T can then be derived by assigning fractional relevancies to documents based on their rank weight in S, and then using these relevancies to calculate the effectiveness metric on T. Such a measure would be symmetric, and seems likely to possess some of the properties sought in an indefinite rank similarity measure, provided that rank weights are chosen so as to create a convergent measure (DCG's weights of $w_i = 1/log(i+1)$ do not). The need for properties such as convergence, and the need to ensure sensible behaviour in limiting cases, means that developing an approach of this kind is not straightforward, and is an area for future investigation. How such an approach would ultimately compare with RBO as it is defined here is not clear.

4. RANK-BIASED OVERLAP

The previous section has shown that the rank similarity measures described in the literature do not meet all the criteria we have identified for similarity measures on indefinite rankings. We now propose a new measure which does meet these criteria: rank-biased overlap (RBO). This is an overlap-based metric, superficially similar to average overlap. The key insight behind RBO is to bias the proportional overlap at each depth by a convergent series of weights (that is, a series whose sum is bounded). As a result, the infinite tail does not dominate the finite head. Therefore, similarity assessment using RBO consists of using prefix evaluation to set upper and lower bounds (Section 4.2) on the score which full evaluation (that is, comparison to infinite depth) could achieve (Section 4.1). In Section 4.3 we derive the weight of each rank under RBO, and therefore the weight of the prefix. The precise full RBO score is, of course, not knowable without evaluation to infinite depth; however, in situations where a single value is needed, a reasonable point estimate can be extrapolated (Section 4.4). Because RBO is a similarity, not a distance, measure, it is not a metric; however, $1 -$ RBO is a metric, as we prove in Section 4.5. Finally, Section 4.6 considers the handling of ties and of rankings of different lengths.

4.1 RBO on infinite lists

We begin by laying out some notation. Let S and T be two infinite rankings, and let S_i be the element at rank i in list S. Denote the set of the elements from position c to position d in list S, that is $\{S_i : c \leq i \leq d\}$, as $S_{c:d}$. Let $S_{:d}$ be equivalent to $S_{1:d}$, and $S_{c:}$ be equivalent to $S_{c:\infty}$. At each depth d, the *intersection* of lists S and T to depth d is:

$$I_{S,T,d} = S_{:d} \cap T_{:d} \ . \tag{1}$$

The size of this intersection is the *overlap* of lists S and T to depth d,

$$X_{S,T,d} = |I_{S,T,d}| \ , \tag{2}$$

and the proportion of S and T that are overlapped at depth d is their *agreement*,

$$A_{S,T,d} = \frac{X_{S,T,d}}{d} \ . \tag{3}$$

For brevity, we will refer to I_d, X_d, and A_d when it is unambiguous which lists are being compared. Using this notation, average overlap can be defined as:

$$\mathrm{AO}(S,T,k) = \frac{1}{k}\sum_{d=1}^{k} A_d \tag{4}$$

where k is the evaluation depth. An example calculation has already been shown in Figure 5.

Consider the family of overlap-based rank similarity measures of the form:

$$\mathrm{SIM}(S,T,w) = \sum_{d=1}^{\infty} w_d \cdot A_d \tag{5}$$

where w is a vector of weights, and w_d is the weight at position d. Then $0 \le \mathrm{SIM} \le \sum_d w_d$, and if w is convergent, each A_d has a fixed contribution $w_d / \sum_d w_d$ (if w is not convergent, then the denominator of this expression goes to infinity). One such convergent series is the geometric progression, where the dth term has the value p^{d-1}, for $0 < p < 1$, and the infinite sum is:

$$\sum_{d=1}^{\infty} p^{d-1} = \frac{1}{1-p} \tag{6}$$

Setting w_d to $(1-p)\cdot p^{d-1}$, so that $\sum_d w_d = 1$, derives rank-biased overlap:

$$\mathrm{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d \ . \tag{7}$$

Rank-biased overlap falls in the range $[0,1]$, where 0 means disjoint, and 1 means identical. The parameter p determines how steep the decline in weights is: the smaller p, the more top-weighted the metric is. In the limit, when $p = 0$, only the top-ranked item is considered, and the RBO score is either zero or one. On the other hand, as p approaches arbitrarily close to 1, the weights become arbitrarily flat, and the evaluation becomes arbitrarily deep.

Rank-biased overlap has an attractive interpretation as a probabilistic user model. Consider a user comparing the two lists. Assume they always look at the first item in each list. At each depth down the two lists, they have probability p of continuing to the next rank, and conversely probability $1-p$ of deciding to stop. Thus, the parameter p models the user's *persistence*. A similar user model was introduced for the retrieval effectiveness metric *rank-biased precision* [Moffat and Zobel 2008]. Once the user has run out of patience at depth d, they then calculate the agreement between the two lists at that depth, and take this as their measure of similarity between the lists. Let D be the random variable giving the depth that the user stops at, and $P(D=d)$ be the probability that the user stops at any given depth d. The expected value of this random experiment is then:

$$\mathbb{E}[A_D] = \sum_{d=1}^{\infty} P(D=d)\cdot A_d \ . \tag{8}$$

Since $P(D = d) = (1 - p) \cdot p^{d-1}$, it follows that $\mathbb{E}[A_D] = \mathrm{RBO}(S, T, p)$. Indeed, this probabilistic model can be extended further by observing that A_d itself gives the probability that an element randomly selected from one prefix will appear in the other. Such probabilistic models help to interpret the meaning of the similarity scores achieved.

4.2 Bounding RBO from prefix evaluation

Rank-biased overlap is defined on infinite lists. Because it is convergent, the evaluation of a prefix sets a minimum and a maximum on the full score, with the range between them being the residual uncertainty attendant upon prefix, rather than full, evaluation. In this section, formulae for the minimum score, $\mathrm{RBO}_{\mathrm{MIN}}$, and the residual, $\mathrm{RBO}_{\mathrm{RES}}$, are derived.

Simply calculating Equation 7 to prefix depth k (let us call this RBO@k) sets a lower bound on the full evaluation, but not a tight one. Indeed, if RBO@k > 0, it is certain that RBO $>$ RBO@k. This is because the overlap in the prefix also contributes to all overlaps at greater depths. (The same problem was observed with average overlap in Figure 5.) More formally, for all $d > k$, $I_d \supseteq I_k$, meaning $X_d \geq X_k$ and A_d is at least X_k/d. Thus, even if all items beyond the prefix turned out on full evaluation to be disjoint, the sum of the agreements at depths beyond k would be:

$$(1 - p) \sum_{d=k+1}^{\infty} \frac{X_k}{d} \cdot p^{d-1} \,. \tag{9}$$

To set a true minimum on full evaluation, Equation 9 is added to the RBO@k score. The infinite sum can be resolved to finite form by the useful equality:

$$\sum_{i=1}^{\infty} \frac{p^i}{i} = \ln \frac{1}{1-p} \,,\ 0 < p < 1 \tag{10}$$

which is derived by integrating both sides of Equation 6. After some rearrangement, we arrive at:

$$\mathrm{RBO}_{\mathrm{MIN}}(S, T, p, k) = \frac{1-p}{p} \left(\sum_{d=1}^{k} (X_d - X_k) \cdot \frac{p^d}{d} - X_k \ln(1-p) \right) \tag{11}$$

where k is the length of the prefix. The $\mathrm{RBO}_{\mathrm{MIN}}(S, T, p, k)$ value gives a tight lower bound on the full $\mathrm{RBO}(S, T, p)$ score. It follows from this that $\mathrm{RBO}_{\mathrm{MIN}}(S, T, p, k)$ is monotonically non-decreasing on deeper evaluation; that is,

$$\forall j > 0,\ \mathrm{RBO}_{\mathrm{MIN}}(S, T, p, j+1) \geq \mathrm{RBO}_{\mathrm{MIN}}(S, T, p, j) \,. \tag{12}$$

Prefix evaluation can also be used to derive a tight maximum on the full RBO score; the residual uncertainty of the evaluation is then the distance between the minimum and maximum scores. The maximum score occurs when every element past prefix depth k in each list matches an element in the other list, beginning with those elements in the prefix that were previously unmatched. Figure 6 illustrates this with an example. The prefix length is $k = 3$, and the overlap X_k at this depth is 1. At each successive depth, two more elements are added, one to each ranking. Therefore, the maximum overlap increases by two until agreement is complete, which occurs at depth $f = 2k - X_k$. Beyond that depth,

d	$S_{:d}$	$T_{:d}$	$\min(A_d)$	$\max(A_d)$	weight
1	<a>	<c>	$0/1$	$0/1$	p^0
2	<ab>	<cb>	$1/2$	$1/2$	p^1
3	<abd>	<cbe>	$1/3$	$1/3$	p^2
4	<abd?[c]>	<cbe?[a]>	$1/4$	$3/4$	p^3
5	<abd??[ce]>	<cbe??[ad]>	$1/5$	$5/5$	p^4
6	<abd???[cef]>	<cbe???[adf]>	$1/6$	$6/6$	p^5
⋮	⋮	⋮	⋮	⋮	⋮
d	<abd...>	<cbe...>	$1/d$	d/d	p^{d-1}

Fig. 6. Minimum and maximum agreements between two indefinite lists at different depths, with evaluation finishing at depth 3. Unseen items for ranks 4 through d are marked as ?. Example hypothetical maximally agreeing elements for these ranks are shown in square brackets.

agreement is fixed at 1. The residual RBO value is therefore:

$$\mathrm{RBO}_{\mathrm{RES}}(S,T,p,k) = (1-p)\left(\sum_{d=k+1}^{f}\frac{2(d-k)}{d}p^{d-1} + \sum_{d=f+1}^{\infty}\left(1-\frac{X_k}{d}\right)p^{d-1}\right)\ . \tag{13}$$

Some rearranging, and again using Equation 10 to reduce the infinite sum, gives:

$$\mathrm{RBO}_{\mathrm{RES}}(S,T,p,k) = p^f + \frac{1-p}{p}\left\{2\sum_{d=k+1}^{f}\frac{(d-k)p^d}{d} - X_k\left[\ln\frac{1}{1-p} - \sum_{d=1}^{f}\frac{p^d}{d}\right]\right\}\ . \tag{14}$$

One might prefer the residual uncertainty of prefix evaluation to be dependent only on the prefix length, not on prefix content. This is not the case with RBO, as prefix agreement determines how long it takes before the difference between the maximum and minimum agreements at subsequent depths d reaches the stationary value of $1 - X_k/d$, as well as this stationary value itself. It is possible, though, to set a range on the values that $\mathrm{RBO}_{\mathrm{RES}}$ can take for a given prefix length, irrespective of prefix contents. The residual will be smallest when $X_k = k$, that is, when the prefix is conjoint. In this case, Equation 13 becomes:

$$\mathrm{RBO}^{\min}_{\mathrm{RES}}(*,*,p,k) \;=\; (1-p)\sum_{d=k+1}^{\infty}\left(1-\frac{k}{d}\right)p^{d-1} \tag{15}$$

$$\;=\; p^k - k\cdot\frac{1-p}{p}\cdot\left(\ln\frac{1}{1-p} - \sum_{d=1}^{k}\frac{p^d}{d}\right)\ . \tag{16}$$

The residual will be largest when $X_k = 0$, that is, when the prefix is disjoint. Then, we have:

$$\mathrm{RBO}^{\max}_{\mathrm{RES}}(*,*,p,k) \;=\; (1-p)\left(\sum_{d=k+1}^{2k}\frac{2(d-k)}{d}\cdot p^{d-1} + \sum_{d=2k+1}^{\infty}p^{d-1}\right) \tag{17}$$

$$\;=\; 2p^k - p^{2k} - 2k\cdot\frac{1-p}{p}\cdot\sum_{d=k+1}^{2k}\frac{p^d}{d}\ . \tag{18}$$

It also follows that $\mathrm{RBO}^{\min}_{\mathrm{RES}}$ will occur when $\mathrm{RBO}_{\mathrm{MAX}} = 1$, and $\mathrm{RBO}^{\max}_{\mathrm{RES}}$ will occur when

$\text{RBO}_{\text{MIN}} = 0$. These formulae are useful in experimental planning. For instance, if two search engines are to be compared on multiple queries, then a first-page or ten-result evaluation with $p = 0.9$ will give a maximum residual of 0.254, for a range of 0.000 to 0.254, and a minimum residual of 0.144, for a range of 0.856 to 1.000. These residuals can be decreased either by examining more results or by using a lower value of p.

Prefix evaluation, then, can be used to set tight bounds upon the full RBO score, meeting our main criteria for a similarity measure on indefinite rankings. The upper and lower limits are monotonically non-increasing and non-decreasing respectively as evaluation continues further down the two lists, in the manner illustrated in Figure 3. Also, RBO_{RES} is monotonically decreasing with evaluation depth: the greater the information about the two lists, the smaller the degree of uncertainty about their full similarity. These monotonic properties are what qualifies RBO to be a satisfactory similarity measure on indefinite rankings. Because of them, the RBO measure provides consistent values for whatever evaluation depth k happens to be chosen, and maintains consistency as this evaluation depth increases. Moreover, the score at any depth of partial evaluation gives strict limits on the score that would be achieved by full evaluation. In contrast, top-k measures are measures only on the lists to depth k, and provide no bounds on the value of full evaluation. Even with partial evaluation, RBO is a measure on the full lists.

4.3 Rank weights under RBO

The agreement at each depth d under RBO is assigned a weight. This weight, however, is not the same as the weight that the elements at rank d themselves take, as these elements contribute to multiple agreements. In this section, we derive a formula for the weight of each rank under RBO. From this, the weight of a prefix can be calculated, which in turn helps guide the choice of the p parameter in the RBO evaluation

The pair of elements at depth d makes no contribution to partial agreements prior to d, takes up $1/d$ th of A_d, $1/(d+1)$ th of A_{d+1}, and so forth. Their precise contribution to the overall score depends on which depth, if any, they are matched at. Consider the difference in the final score between, on the one hand, both elements at depth d being matched at or prior to depth d (maximum agreement), and, on the other, neither element being matched at infinite depth (minimum agreement). We will refer to this difference as the *weight* of rank d, denoted as $W_{\text{RBO}}(d)$. Accounting for the weighting of the agreements $w_d = (1-p) \cdot p^{d-1}$ (Equation 7), the weight of rank d under RBO is therefore:

$$W_{\text{RBO}}(d) = \frac{1-p}{p} \sum_{i=d}^{\infty} \frac{p^i}{i} \tag{19}$$

The weight of the prefix of length d, $W_{\text{RBO}}(1:d)$, is then the sum of the weights of the ranks to that depth:

$$W_{\text{RBO}}(1:d) = \sum_{j=1}^{d} W_{\text{RBO}}(d) = \frac{1-p}{p} \sum_{j=1}^{d} \sum_{i=j}^{\infty} \frac{p^i}{i} \tag{20}$$

which after some rearrangement, and using Equation 10 to resolve the infinite sum, gives:

$$W_{\text{RBO}}(1:d) = 1 - p^{d-1} + \frac{1-p}{p} \cdot d \cdot \left(\ln \frac{1}{1-p} - \sum_{i=1}^{d-1} \frac{p^i}{i} \right) . \tag{21}$$

Of course, the weight of the tail, $W_{\mathrm{RBO}}(d+1:\infty)$, is $1 - W_{\mathrm{RBO}}(1:d)$. And since $W_{\mathrm{RBO}}(1:d)$ is invariant on the length of the list, it follows that the weight of the infinite tail does not dominate that of the finite head.

Equation 21 helps inform the choice of the parameter p, which determines the degree of top-weightedness of the RBO metric. For instance, $p = 0.9$ means that the first 10 ranks have 86% of the weight of the evaluation; to give the top 50 ranks the same weight involves taking $p = 0.98$ as the setting. Thus, the experimenter can tune the metric to achieve a given weight for a certain length of prefix.

4.4 Extrapolation

Definitions of $\mathrm{RBO}_{\mathrm{MIN}}$ and $\mathrm{RBO}_{\mathrm{RES}}$ have been formulated in Section 4.2. The RBO score can then be quoted either as base+residual or as a min–max range. For many practical and statistical applications, though, it is desirable or necessary to have a single score or point estimate, rather than a range of values.

The simplest method is to use the base RBO value as the single score for the partial evaluation. The base score gives the known similarity between the two lists, the most that can be said with certainty given the information available. However, the base score is dependent on the evaluation depth, k. The highest base score that can be achieved for depth k evaluation using persistence p is:

$$1 - p^k - \frac{k(1-p)}{p}\left(\sum_{d=1}^{k}\frac{p^d}{d} + \ln(1-p)\right) \tag{22}$$

which, for large p and small k, is well short of 1. There are practical situations in which a list is conceptually indefinite but where only the first few items are available. For instance, if two search engines each only supply 7 results to a query, and the p parameter employed is 0.9, then even if both results lists are identical (to the supplied depth), the base RBO score will only be 0.767. In such situations, base RBO can easily become a measure of result list length, not difference.

An alternative formulation for a single RBO score is to extrapolate from the visible lists, assuming that the degree of agreement seen up to depth k is continued indefinitely. Denote as $\mathrm{RBO}_{\mathrm{EXT}}$ the result of such an extrapolation. To derive a direct formula for $\mathrm{RBO}_{\mathrm{EXT}}$, we start from Equation 9, which gives the adjustment to the RBO value, calculated on the k seen items, to make it a true minimum value. The assumption for the lower bound is that the remaining items are all non-conjoint, so that the agreement at ranks $r > k$ is X_k/r. Instead, extrapolation assumes that the degree of agreement seen at k is expected to continue to higher ranks, that is, that for $r > k$, $A_r = X_k/k$. (The resulting agreement values may not in reality be possible, because they would require fractional overlap. Consider, though, the analogy of the expected value of a random experiment not having to be a possible outcome of that experiment; for instance, the expected value of rolling a fair six-sided die is 3.5.) Constant agreement considerably simplifies things, resulting in:

$$\mathrm{RBO}_{\mathrm{EXT}}(S,T,p,k) = \frac{X_k}{k}\cdot p^k + \frac{1-p}{p}\sum_{d=1}^{k}\frac{X_d}{d}\cdot p^d\,. \tag{23}$$

It should be noted that this is not equivalent to simply extrapolating a score between the numeric values of $\mathrm{RBO}_{\mathrm{MIN}}$ and $\mathrm{RBO}_{\mathrm{MAX}}$. Since those scores are weighted to higher ranks, such an extrapolation would also be weighted to the agreement observed in higher ranks.

Instead, RBO_{EXT} extrapolates out from A_k, that is, the agreement observed at evaluation depth k.

Extrapolated RBO is not monotonic; it could either increase or decrease as the prefix lengthens. However, RBO_{EXT} will always increase with increasing agreement and decrease with decreasing agreement. That is, if $A_{d+1} > A_d$ then $\text{RBO}_{\text{EXT}}(d+1) > \text{RBO}_{\text{EXT}}(d)$, and conversely if $A_{d+1} < A_d$ then $\text{RBO}_{\text{EXT}}(d+1) < \text{RBO}_{\text{EXT}}(d)$, for all $d > 0$. It was noted in Section 3.4 that this property is not observed by average overlap. And of course, RBO_{EXT} is bounded, by RBO_{MIN} and RBO_{MAX}.

Where a point score is needed, there is the choice of RBO_{BASE} or RBO_{EXT}. In many cases, evaluation will be performed deeply enough, and p will be small enough (say, $p \leq 0.9$ and depth of 50), that the residual disappears at normal reporting fidelity, leaving RBO_{EXT} and RBO_{BASE} as indistinguishable and almost-exact estimates of the true RBO score. Where the residual is noticeable, RBO_{EXT} should in general be the preferred point estimate, in part because it is less sensitive than RBO_{BASE} to the actual evaluation depth, which may vary between different ranking pairs in the one experiment. For noticeable residuals, the full reporting format is $\text{RBO}_{\text{EXT}}[\text{RBO}_{\text{MIN}} - \text{RBO}_{\text{MAX}}]$.

4.5 Metricity

Since RBO measures similarity, not distance, it is not a metric. However, RBO can be trivially turned into a distance measure, rank-biased distance (RBD), by $\text{RBD} = 1 - \text{RBO}$. We now prove that RBD is a metric.

PROPOSITION 4.1. *RBD is a metric.*

Proof. Since RBO is clearly symmetric, it is sufficient to show that the triangle inequality holds, that is,

$$\forall R, S, T,\ \text{RBD}(R, T, p) \leq \text{RBD}(R, S, p) + \text{RBD}(S, T, p) \cdot \tag{24}$$

Now

$$\begin{aligned} \text{RBD}(S, T, p) &= 1 - \text{RBO}(S, T, p) \\ &= 1 - (1-p) \sum_{d=1}^{\infty} \frac{|S_{:d} \cap T_{:d}|}{d} \cdot p^{d-1} \\ &= (1-p) \sum_{d=1}^{\infty} \frac{|S_{:d} \triangle T_{:d}|}{2d} \cdot p^{d-1} \end{aligned} \tag{25}$$

where $\triangle$ is symmetric difference, that is, the elements that are in one set or the other but not both. The last simplification is derived from the fact that:

$$\begin{aligned} 2d = |S_{:d}| + |T_{:d}| &= |S_{:d} \triangle T_{:d}| + 2 \cdot |S_{:d} \cap T_{:d}| \\ \Rightarrow 1 - \frac{|S_{:d} \cap T_{:d}|}{d} &= \frac{|S_{:d} \triangle T_{:d}|}{2d} \end{aligned} \tag{26}$$

So $\text{RBD}(S, T)$ is the weighted sum of these $|S_{:d} \triangle T_{:d}|$, where the weighting is invariant on the contents of the list. Therefore, we need only demonstrate that

$$\forall d, |R_{:d} \triangle T_{:d}| \leq |R_{:d} \triangle S_{:d}| + |S_{:d} \triangle T_{:d}| \tag{27}$$

The remainder of the proof follows Fagin et al. [2003]. Consider an element $x \in R \triangle T$. Assume, without loss of generality, that $x \in R$; therefore, $x \notin T$. There are two cases:

$x \in S$, in which case $x \in S \triangle T$ but $x \notin R \triangle S$; or $x \notin S$, in which case $x \in R \triangle S$ but $x \notin S \triangle T$. Either way, if an element occurs on (contributes to) the left side of Equation 27, it must occur on (contribute to) the right side. Equation 27 then holds, and therefore so does Equation 24. □

Similar proofs hold for the metricity of $1 - \mathrm{RBO}_{\mathrm{MIN}}$ and $1 - \mathrm{RBO}_{\mathrm{EXT}}$.

4.6 Ties and uneven rankings

Ties may be handled by assuming that, if t items are tied for ranks d to $d + (t-1)$, they all occur at rank d. To support this, we modify the definition of agreement given in Equation 3:

$$A_{S,T,d} = \frac{2 \cdot X_{S,T,d}}{|S_{:d}| + |T_{:d}|} \,. \tag{28}$$

Equations 3 and 28 are equivalent in the absence of ties extending over rank d, but in the presence of such ties, the former formulation can lead to agreements greater than 1.

It occasionally happens that indefinite rankings are compared with different evaluation depths on each ranking. One cause of such irregularity is that the providers of the rankings are returning lists shorter than the evaluation depth chosen for the assessment and different from each other. We will call such lists *uneven rankings*. For instance, for an obscure but not entirely nonsensical query, one public search engine might return five results, another might return seven. These can still be treated as indefinite rankings; there are many more web pages beyond these depths, but they have not met the engine's threshold of estimated relevance. For the following discussion, let L be the longer of the two lists, with length l, and S be the shorter, with length s.

The formula for $\mathrm{RBO}_{\mathrm{MIN}}$ given in Equation 11 handles uneven rankings without modification, since it is implicitly assumed that $\forall d \in \{s+1, \ldots, l\}, S_d \notin L$; that is, we assume maximal disjointness and are done with it. Conversely, $\mathrm{RBO}_{\mathrm{MAX}}$ is found by assuming that every item in the extension of S matches one item in L, increasing the overlap by one. Therefore, $\forall d \in \{s+1, \ldots, l\}, X_d^{\max} - X_d^{\min} = d - s$, regardless of the contents of the preceding lists. The definition of $\mathrm{RBO}_{\mathrm{RES}}$ on uneven lists then becomes:

$$\mathrm{RBO}_{\mathrm{RES}}(L,S,l,s) = \\ \frac{1-p}{p}\left(\sum_{d=s+1}^{l}\frac{d-s}{d}p^d + \sum_{d=l+1}^{f}\frac{2d-l-s}{d}p^d + \sum_{d=f+1}^{\infty}\left(1-\frac{X_l}{d}\right)p^d\right) \tag{29}$$

where $f = l + s - X_l$ is the rank at which maximum agreement becomes 1. Removing the infinite sum using Equation 10 once again, and simplifying, results in:

$$\mathrm{RBO}_{\mathrm{RES}}(L,S,l,s) = \\ p^s + p^l - p^f - \frac{1-p}{p}\left(s\sum_{d=s+1}^{f}\frac{p^d}{d} + l\sum_{d=l+1}^{f}\frac{p^d}{d} + X_l\left[\ln\frac{1}{1-p} - \sum_{d=1}^{f}\frac{p^d}{d}\right]\right) \tag{30}$$

Modifying $\mathrm{RBO}_{\mathrm{EXT}}$ to handle uneven rankings is less straightforward. The extrapolation for even rankings is achieved by assuming the agreement in the unseen part of the lists is the same as in the prefixes. However, agreement between L and S is not known to depth l. And while agreement to depth s is known, truncation at this depth loses information on the degree of overlap between $L_{(s+1):l}$ and S. Therefore, extrapolation for uneven rankings must separately extrapolate agreement for $S_{(s+1):l}$.

Consider the method of extrapolation for even lists. The agreement A_k at common evaluation depth k is assumed to continue unchanged at further evaluation depths. In other words, $\forall d > k$, $A_d = A_k$, and specifically $A_{k+1} = A_k$. Referring to the definition of agreement in Equation 3, this means that

$$|S_{:k+1} \cap T_{:k+1}| \stackrel{\text{def}}{=} X_{k+1} = X_k + A_k \ . \tag{31}$$

If $0 < A_k < 1$, which is generally the case, then working backwards through the formula implicitly requires $X_{d>k}$ to take on fractional values. This suggests the concept of degree of set membership. An element occurring in the seen prefix will have a membership degree of 1 or 0, depending on whether it is matched in the other list at the current evaluation depth. An unseen element, however, is assigned under extrapolation a (usually fractional) membership degree; one could think of it as a "probability of membership". The elements S_{k+1} and T_{k+1} in Equation 31, for even lists, each have membership A_k. In the case of uneven lists, the conjointness of $L_{(s+1):l}$ is known to be either 0 or 1. Nevertheless, the membership of the unseen elements $S_{(s+1):l}$ can still be set to A_s. This will provide an assumed A_l, which can be extrapolated for elements beyond depth l, unseen in both lists. The formula then is:

$$\mathrm{RBO}_{\mathrm{EXT}}(L, S, l, s) = \frac{1-p}{p}\left(\sum_{d=1}^{l}\frac{X_d}{d}p^d + \sum_{d=s+1}^{l}\frac{X_s(d-s)}{sd}p^d\right) + \left(\frac{X_l - X_s}{l} + \frac{X_s}{s}\right)p^l \tag{32}$$

Note that X_l here means the overlap on the seen lists at depth l, even though $|S| < l$; the maximum value of X_l is therefore s.

Calculating $\mathrm{RBO}_{\mathrm{EXT}}$ on uneven lists in this way maintains two important criteria met by extrapolation on even lists. First, $\mathrm{RBO}_{\mathrm{MIN}} \leq \mathrm{RBO}_{\mathrm{EXT}} \leq \mathrm{RBO}_{\mathrm{MAX}}$. And second, $\mathrm{RBO}_{\mathrm{EXT}}$ is non-increasing with deeper evaluation if S_{s+1} or L_{l+1} is found to be disjoint, and non-decreasing if the element is found to be conjoint.

5. EXPERIMENTAL DEMONSTRATIONS

Section 4 has defined the RBO metric, and described how it meets the criteria for an indefinite rank similarity measure, which the measures discussed in Section 3 failed to do. We now illustrate the use of RBO, first in comparing document rankings produced by public search engines, and secondly as an experimental tool in the research laboratory of the IR system developer. These domains involve non-conjoint rankings, so rank similarity measures such as τ that require conjointness cannot be applied. The only viable alternatives to RBO are other non-conjoint rank similarity measures. We provide comparisons with two of these: Kendall's distance (KD) and average overlap (AO).

5.1 Comparing search engines

We begin by using RBO to compare the results returned by public search engines. Twenty search engine users, drawn from the authors' colleagues and acquaintances, were asked to provide search queries taken either from their recent search history or as examples of queries the might currently be searching for. Each user returned between three and eight queries, making a total of 113 queries (available from the authors on request), collected from mid-August to early September 2008. The queries were then submitted once a day to a number of public search engines, beginning on October 20th, 2008, and running up until

Name	URL	Notes
Google	`www.google.com`	Global Google. Google maps, news, blog results excluded; Google books results retained.
Yahoo	`search.yahoo.com`	Global Yahoo!.
Live	`search.live.com`	Global Live.
Ask	`www.ask.com`	Ask.
Dogpile	`www.dogpile.com`	Dogpile. Maximum 80 results.
Sensis	`www.sensis.com.au`	Sensis. Maximum 10 results. Not restricted to Australian-only results.
Alexa	`www.alexa.com`	Alexa.
A9	`a9.com`	A9. Maximum 20 results. Ceased offering general web search in January 2009. Prior to that, results based on Alexa.
Google (AU)	`www.google.com.au`	Google Australia search portal. Not restricted to Australian-only results.
Yahoo (AU)	`au.search.yahoo.com`	Yahoo! Australia search portal. Not restricted to Australian-only results.
Live (AU)	`www.live.com`	Specified `\?mkt=en-au`. Not restricted to Australian-only results.

Table II. Public search engines searched.

February 26th, 2009. Eleven different search engines were searched, as listed in Table II. Three of these are the Australian portals of international search engines. In every case, queries were submitted directly to the web site via URL manipulation and the HTML results list was scraped; the search APIs of these engines were not used. Except where noted, the top 100 results were retrieved from each search engine. In a given results list, only the first result from any given host was retained; most search engines only provided a maximum of two results from the one host, with the second being folded directly under the first. Result URLs were captured as returned by the search engines; no further normalization was performed.

Public search engines commonly return ten search results per page, including on the first results page. Therefore a reasonable choice of the p parameter is one that sets the expected number of results compared by the p-persistent user to 10. This is achieved by setting p to 0.9. As described in Section 4.3, this is equivalent to giving the first ten results 86% of the weight in the similarity comparison. It is also convenient to concentrate on the top ten results because, for interface reasons, it was not practical to retrieve more than the first ten results from some search engines. This again illustrates the importance of a rank similarity measure being monotonic in depth: we will be comparing rankings with a variety of depths, some going to depth 100, others to depth 10, and yet others somewhere in between, and we want the similarity scores produced to be comparable across all cases.

Table III gives the mean RBO_{EXT}, $p = 0.9$, between the different global search engines across all 113 queries on December 5th, 2008. The key to interpreting the numerical value of these scores is to remember that RBO is a measurement of expected overlap, or equivalently of a weighted mean of overlap at different depths. Thus, the RBO score of 0.25 between Google and Live can very roughly be understood as saying that the two systems have 25% of their results in common (as a decaying average over increasing depths). Contrary perhaps to expectations, different search engines are in fact returning quite different results, or at least result orderings, to queries; only a handful have an RBO above 0.25. By

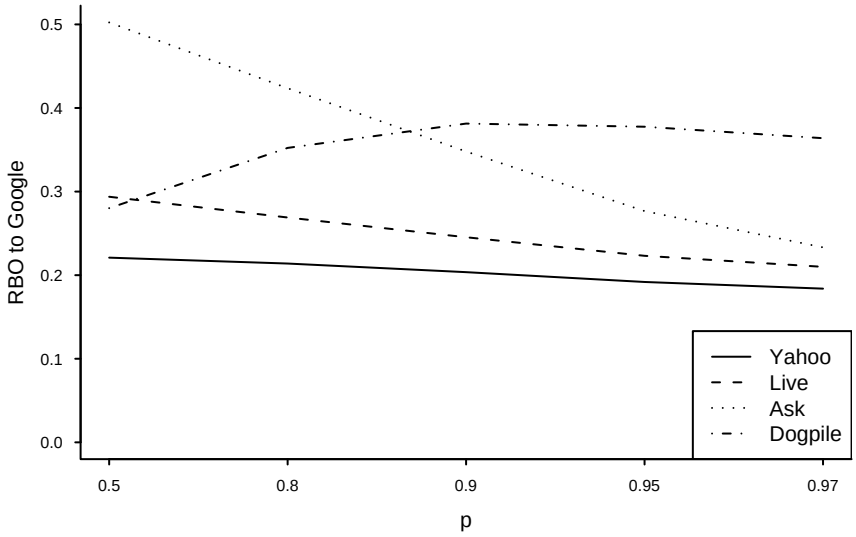


Fig. 7. Mean RBO between Google and other search engines for different values of the p parameter. Raising p increases the depth of comparison.

the date of this run of queries, Alexa had started to draw its results from Live, which is why their RBO score is so high. Previously, Alexa had been an independent search engine, which A9 drew its general search results from, and these two engines had a very high RBO (around 0.9). Not long after December 5th, 2008, A9 stopped offering general-purpose web search and became solely a product search aggregator. The Dogpile engine aggregates results from Google, Yahoo, Live, and Ask. The RBO figures suggest that Google results are given the strongest weighting by Dogpile; the fact that Ask is higher than Yahoo and Live may be because Ask is itself closer to Google. The Sensis search engine is quite unlike all the others, as to a lesser extent is A9. Table IV shows the RBO between the global and Australian-localized search results for the search engines that provide localized variants. Apparently, Google performs much lighter localization than either Yahoo or Live.

Other values than 0.9 could reasonably be chosen for the p parameter in search engine comparisons. The researcher might wish to concentrate more heavily on the user experience of the first few results, in which case p values of 0.8 or even 0.5 might be appropriate,

	yahoo	live	ask	dogpile	sensis	alexa	a9
google	0.20	0.25	0.35	0.38	0.03	0.23	0.11
yahoo		0.21	0.17	0.24	0.03	0.21	0.08
live			0.18	0.24	0.03	0.76	0.10
ask				0.27	0.04	0.17	0.09
dogpile					0.03	0.23	0.08
sensis						0.03	0.02
alexa							0.09

Table III. Mean RBO, $p = 0.9$, between non-localized search engines across 113 user queries issued on 2008-12-05.

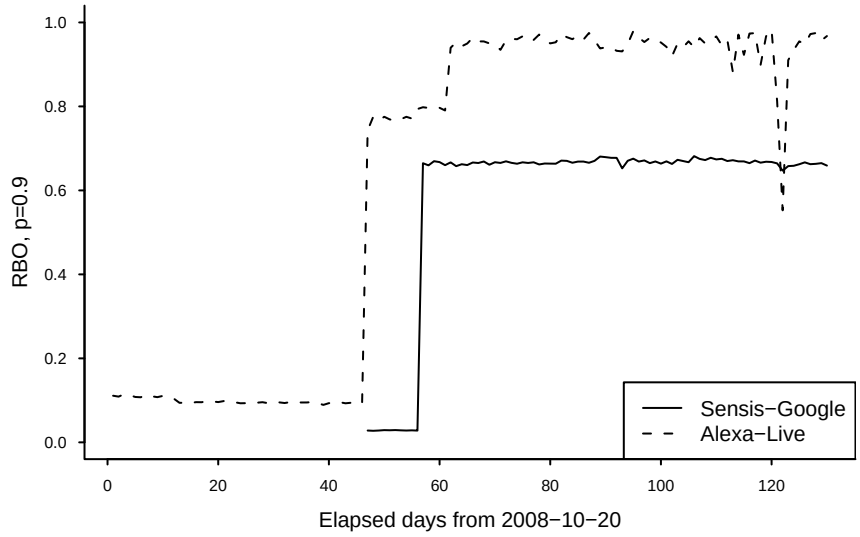


Fig. 8. Mean RBO, $p = 0.9$, across 113 queries between Sensis and Google, and between Alexa and Live, calculated daily over the experimental period.

leading to expected comparison depths of 5 and 2, respectively. Conversely, a deeper, more system-centric comparison might be preferred, suggesting p values of 0.95 or 0.97 (expected depths of 20 and 33.3). Or the researcher might be interested to contrast a range of comparison depths. Because RBO's top-weightedness is tuneable via the p parameter, such investigations are possible. A question that can be addressed in this way is whether search engines are more similar to each other at the top of their rankings than further down. Raising the p value deepens the comparison, allowing us to explore this hypothesis. Figure 7 shows that Yahoo and Live are indeed more similar to Google at higher ranks than lower, but only mildly so. The difference is much stronger for Ask, suggesting that it is (by design or coincidence) strongly tuned towards delivering a similar first-page experience to Google. The rise, with increasing depth, of Dogpile's similarity to Google in Figure 7 might on a naive reading lead to the (surprising) interpretation that Dogpile draws more results from Google further down the ranking than higher up. But this interpretation fails to appreciate that aggregated results are supplementary, one engine drawing in another's answers; Dogpile's similarity to Live and Yahoo (not shown) rises even more strongly with depth. The function of RBO here is to alert us to an anomaly: Google's relationship to Dogpile is quite different from its relationship to the other engines.

During the period of the study, Sensis ceased being an independent search engine, and switched to deriving its results from Google. Similarly, Alexa changed to deriving its results from Live. These events can be traced by looking at the mean RBO scores of the

	google	yahoo	live
RBO	0.77	0.35	0.44

Table IV. Mean RBO, $p = 0.9$, between localized and non-localized search engines across 113 queries issued on 2008-12-05.

	Day-to-day			Start-to-end
	mean	sd	median	mean
google	0.91	0.08	0.94	0.50
yahoo	0.94	0.09	0.98	0.45
live	0.94	0.12	1.00	0.43
ask	0.94	0.13	1.00	0.41

Table V. Rate of change of search engine results over time, as measured by RBO between sequential daily runs (left) and between start and end of experiment (right).

respective system pairs over time, as displayed in Figure 8. Evidently, Sensis switched to using Google on Day 57 (December 15th, 2008), while Alexa moved to using Live on Day 47 (December 5th, 2008), initially with some modifications, and almost verbatim from Day 62 (December 20th). The dip in similarity between Alexa and Live on Day 122 (February 18th, 2009) is due to Alexa giving idiosyncratic results on this day; why it does so is not clear. (Due to a problem with the query processor, complete results are not available for Sensis prior to Day 47.) Kendall's distance and average overlap detect similar overall trends to those shown in Figure 8, but show relatively greater similarity between Sensis and Google after the switch. We hypothesize that Sensis may be seeding (possibly localized) results into the top of the ranking provided by Google. The top-weightedness of RBO would detect such top-heavy seeding more effectively that Kendall's distance or average overlap.

Another question of interest is how much the results of different search engines change over time. This gives a sense of how dynamic a search service is, either by way of crawling policy, or through changes in its ranking computation. For each of the 113 queries, the RBO between one day's results and the following day's results was calculated, for all 129 days in the experimental set. For each search engine, the mean and median across all day-to-day RBO scores were calculated, as was the mean of the standard deviation of RBO scores for each query over time. The results are shown in Table V. Results tend to be relatively stable from one day to the next; indeed, for Live and Ask, the "typical" (median) result does not change at all. The results from Google show the highest rate of change. Additionally, changes to Google results, and to a lesser extent those of Yahoo, tend to be constant and even (median closer to mean, low standard deviation), whereas changes to Live and Ask results are more sporadic (median further from mean, high standard deviation). Also shown is the mean RBO between result lists taken from towards the beginning of the experiment (Days 16 through 19) and then towards the end (Days 111 through 114), 16 pairs in total for each query and each system. Query results have shifted significantly over the three months, but systems are still more similar to the time-shifted versions of themselves than (referring back to Table III) they are to each other. Interestingly, while Google shows more day-to-day change, it shows the least amount of long-term change. This latter result is significant in a two-sample, two-tailed t-test at 0.05 level between Google and each of the other search engines, but differences between the other engines are not significant.

It is informative to compare the RBO results with those obtained by using Kendall's distance at depth $k = 100$, reported in Table VI. The large degree of disjointness between results causes Kendall's distance to return negative values for all except the derivative Live–Alexa pair. Negative values make little sense in this application: there is no sense

in which any of these search engines are giving rankings negatively correlated with any other. Kendall's distance gives different relative results than RBO in a number of cases. For instance, RBO reports Dogpile to be closest to Google, but Kendall's distance places it closer to Yahoo; this is because on average Dogpile appears to pull more results from Yahoo than from Google (mean agreement at 100 is 0.49 for Yahoo, and 0.30 for Google), but seems to give a higher ranking to the results from Google. Similarly, of the independent systems, RBO places Ask as being closest to Google, whereas Kendall's distance places it as being farthest away; again, in this case, Kendall's distance is following agreement at one hundred. Thus, although Kendall's distance is by design a correlation metric, its lack of top-weightedness and the highly non-conjoint nature of these indefinite rankings has it tending towards an unweighted measure of set agreement.

Too much significance should not be attached to these results as they stand. A rigorous examination of search engine similarities would start from these high-level RBO figures, not finish with them. Nevertheless, these comparisons do give a flavour of the analysis that a suitable rank similarity measure allows us to make upon search engine results, and indicate that RBO is uniquely suitable for these purposes.

5.2 Experimenting with information retrieval

In this section, we examine the use of RBO in a typical research situation, where an IR system is being modified, and the researcher wishes to measure how much the modification is changing the results. The researcher may be using the rank similarity measure as a proxy for a retrieval effectiveness metric. For instance, an efficiency change might have been made, and the rank similarity comparison is being used as an indicator of the degradation in effectiveness that the change could have caused, as with our first example below. Using RBO is attractive in this situation because performing the relevance assessments needed for effectiveness evaluation is expensive. If an initial examination with RBO determines that only slight changes have occurred in (top-weighted) ranking order for some or all topics, then the expense of relevance assessment on those topics can be avoided. Or the researcher may simply be measuring ranking fidelity as such, as in our second example.

Query pruning was mentioned in Section 2. It is a technique in which the amount of memory that is used in query processing is limited, and the amount of processing time reduced, but at a possible cost in retrieval accuracy and effectiveness. Therefore, if the results of a pruned system differ from those of an unpruned one, this suggests (though does not by itself prove) a degradation in effectiveness. Figure 9 gives the results of using RBO and Kendall's distance in a query pruning experiment. The query-pruned results are compared to the unpruned results, with evaluation carried out to varying depths. Here

	yahoo	live	ask	dogpile	sensis	alexa	a9
google	−0.60	−0.56	−0.66	−0.20	−0.93	−0.58	−0.80
yahoo		−0.55	−0.75	−0.04	−0.94	−0.56	−0.85
live			−0.73	−0.31	−0.93	0.62	−0.81
ask				−0.41	−0.91	−0.73	−0.83
dogpile					−0.93	−0.35	−0.82
sensis						−0.93	−0.95
alexa							−0.83

Table VI. Mean Kendall's distance at depth 100 between non-localized search engines across 113 user queries issued on 2008-12-05.

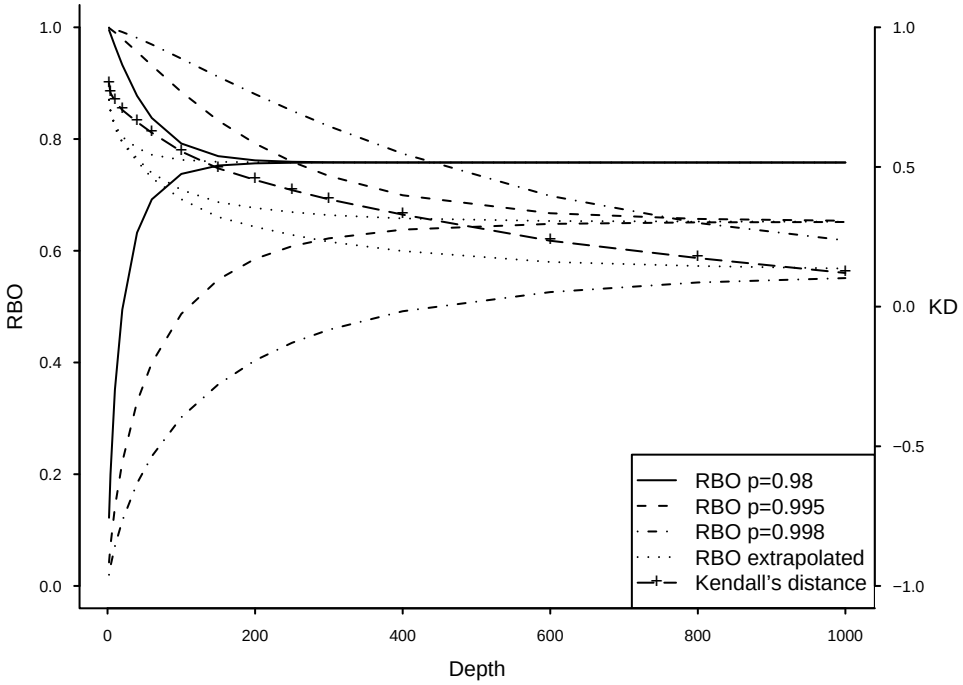


Fig. 9. Similarity of query-pruned and unpruned runs. Kendall's distance and RBO with different p parameters are calculated at increasing depths, averaged across all topics. The upper and lower bounds and extrapolated values are shown for RBO. The corpus is wt18g, and the queries are TREC queries 551–600, title only. The retrieval engine is Zettair $0.9.3$, using the Dirichlet similarity metric. Pruning is as described in Lester et al. [2005], with a limit of $1{,}000$ accumulators, compared to no accumulator limit.

the unpruned results are the objective or "gold-standard" ranking, from which the pruned results deviate. All extrapolated RBO values and also Kendall's distance decrease as the depth of evaluation increases. This is because query pruning tends to have a greater effect on late-ranking than top-ranking documents. The extrapolated RBO value asymptotes to its final value relatively quickly, even for the very deep $p = 0.998$ evaluation. On the other hand, the Kendall's distance score is still falling at depth $1{,}000$, and it is not clear what value it is asymptoting to, if any. We see clearly here that Kendall's distance is a measure, not on the full list, but on the prefix. In contrast, base plus residual RBO is a measure on the full list, and even the extrapolated value shows greater stability. It should be noted that all the p values chosen here are quite high. If one were using RBO as a proxy for a retrieval effectiveness metric, $p = 0.98$ would be at the upper end of the values one would be likely to choose, in which case the value has already converged by depth 200.

Figure 10 shows a different kind of alteration to an information retrieval system. In this case, a language model smoothed with Dirichlet priors is being used to score the similarity between query and documents. This query–document similarity measure takes

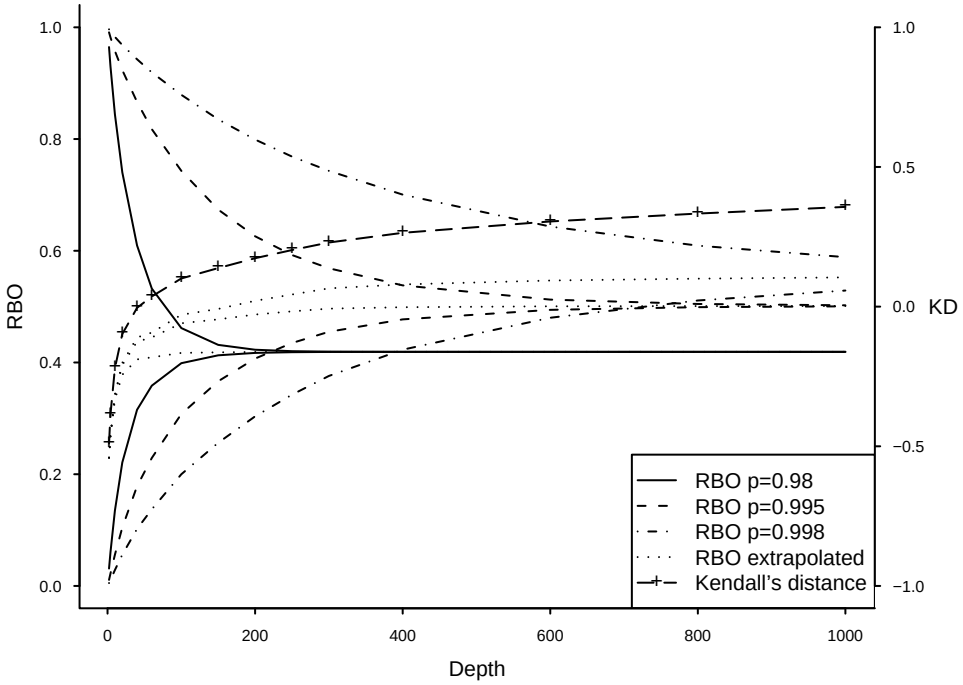


Fig. 10. Similarity of runs with different similarity metric tuning parameters. Kendall's distance and RBO with different p parameters are calculated at increasing depths, averaged across all topics. The upper and lower bounds and extrapolated values are shown for RBO. The corpus is wt18g, and the queries are TREC queries 551–600, title only. The retrieval engine is Zettair $0.9.3$, using the Dirichlet similarity metric. The μ parameter of the Dirichlet metric was set to 500 for one run, and $5{,}000$ for the other.

a parameter μ, which balances the influence of the relative weighting of terms within a document: with lower μ values, relative weighting is emphasized, meaning some terms have significantly higher impact than others, whereas with higher μ values, each term tends to have similar weighting and what matters is simply the presence or absence of the term [Zhai and Lafferty 2004]. Two different values of μ are being tried in Figure 10 as part of a parameter tuning experiment, with the mean RBO across a set of topics being displayed. Here, neither parameter value is the baseline or objective value, from which the other parameter is deviating and presumably degrading. Rather, the interest is in seeing how much of a difference is caused by altering the parameter. In contrast to Figure 9, the RBO_{EXT} and Kendall's distance scores trend up as depth of evaluation increases, not down. The reason is that parameter tuning tends to cause localized perturbations in ordering; as the depth increases, the degree of overlap increases too. All point measures give rising similarity values with depth, but Kendall's distance rises considerably more than even the highest-p RBO, and it appears not to have asymptoted by depth $1{,}000$, even though the extrapolated RBO values stabilize well before that. Even though Kendall's distance is

derived from a metric that is based upon counting perturbations, it seems to be even more strongly affected by overlap than RBO itself is.

Of course, the preceding two cases are only examples. Different ranking perturbations will result in different effects on rank similarity measures. Nevertheless, these examples serve to illustrate two important points. The first is that the values of non-convergent measures evaluated to shallow depths can be very different from those at deeper depths, and so such measures cannot be regarded as adequate similarity measures on indefinite rankings. In contrast, a convergent metric gives hard bounds on infinite evaluation. The second, related point is that Kendall's distance and other top-k metrics cannot be regarded as single measures, but rather as families of measures, with each k value instantiating a different member of the family. That is, Kendall's distance is at least as dependent on its cutoff depth k as RBO is on its parameter p.

5.3 Correlation with effectiveness measures

We conclude by examining the relationship between rank similarity measures and changes in retrieval effectiveness. The metric used to calculate retrieval effectiveness is average precision (AP), which is defined as follows. Let the precision of a ranking to depth k be the proportion of documents to depth k that are relevant. The sum of precisions for that ranking is the sum of the precision at each ranking that a relevant document is returned. Average precision for the ranking is then the sum of precisions divided by the total number of (known) relevant documents for that query. To calculate the correlation between effectiveness and rank similarity measures, one could take actual retrieval runs, perturb their rankings, and calculate the similarity between the original and perturbed rankings on the one hand, and the change in effectiveness on the other. Actual rankings, however, are typically far from ideal ones, so randomly perturbing them, while decreasing the ranking similarity, has a rather noisy influence on effectiveness. Instead, we take a simulated approach. An ideal ranking of 10 relevant and 90 irrelevant documents is progressively degraded. The degradation consists of a sequence of 25 swaps between a relevant and a non-relevant document, chosen at random. After each such swap, the AP of the degraded ranking, and similarity of the degraded to the ideal ranking, is calculated and plotted. For calculating AP, the total number of relevant documents is set to 10 (that is, the retrieval system has retrieved all relevant documents).

The results of this simulated experiment are given in Figure 11. A total of 100 degradations were performed; each of the above figures therefore consists of $2{,}500$ points. The Kendall's τ between the AP score and the similarity value of the data points is also displayed. Kendall's distance shows a weaker correlation with AP than either of the top-weighted metrics. Moreover, it is more sensitive to the cutoff point. Cutoff at 10 gives the best correlation with AP across the whole sequence, but poor correlation at the top, and insensitivity to relationships beyond depth 10. Evaluation to depth 100 shows quite poor correlation. Average overlap shares some of this sensitivity to evaluation depth, whereas RBO has high fidelity at high similarity, regardless of the p value chosen. A comparison between the average overlap and RBO figures illustrates how intimately average overlap is linked with the choice of cutoff depth. Cutoff depth has at least as strong an effect on average overlap as changes in the p parameter has on RBO, even though as argued before cutoff depth is essentially arbitrary in an indefinite ranking.

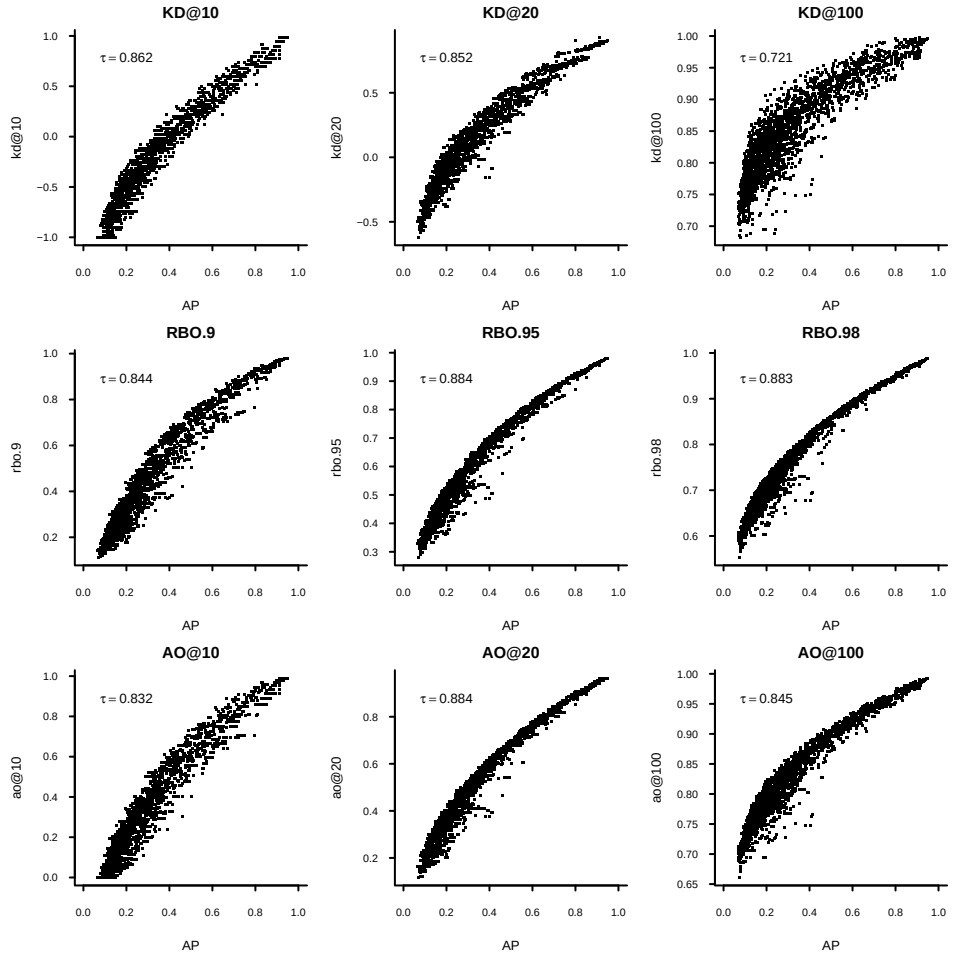


Fig. 11. Correlation between the average precision (AP) of a degraded ranking on the one hand, and rank similarity between the degraded and the ideal ranking on the other, for the experiment in which we start with a ranking of 10 relevant followed by 90 non-relevant documents, then randomly swap relevant and non-relevant elements 25 times, recording similarity and AP at each iteration, with 100 independent repetitions. The similarity metrics used are Kendall's distance (KD) at different depths; rank-biased overlap (RBO) with different p values; and average overlap (AO) at different depths.

6. CONCLUSIONS

Non-conjoint, top-weighted, and incomplete ranked lists – what we have called *indefinite* rankings – are ubiquitous. Appropriate measures of their similarity, however, are lacking. Such a measure must take into account all the peculiar characteristics of indefinite rankings. It must be top-weighted, giving more emphasis to the degree of similarity at the top of the ranking than further down. It must handle non-conjointness in the rankings, neither requiring every item to appear in both rankings, nor making arbitrary assumptions about where items uniquely seen in one ranking are located beyond the prefix in the other. And

finally, it must recognize that the observed rankings are incomplete prefixes of much longer full rankings, and that the cutoff depth of the prefix is essentially arbitrary. A corollary of this incompleteness is that what is desired is a measure of the similarity of the full rankings, not merely of the observed prefixes. No existing similarity measure on ranked lists meets all of the above requirements.

In this paper, we have introduced a new similarity measure on ranked lists, namely rank-biased overlap, or RBO. It is tuneably top-weighted, handles non-conjointness in the rankings, and is not tied to a particular prefix length. Most importantly, it is a similarity measure on the full rankings, even when only a prefix of each is available for comparison. It achieves this by using a convergent set of weights across successive prefixes, preventing the weight of the unseen tail from dominating that of the observed head. As a result, partial evaluation allows us to set strict upper and lower bounds on the similarity of the full rankings – a similarity whose exact value could only be calculated by evaluating the rankings in full. The RBO measure is parameterized to tune the degree of top-weightedness, and we have provided guidelines on the parameter choice. An extrapolated RBO value has been derived to give a reasonable point estimate on this similarity. This extrapolated value is itself monotonic on agreement. If the degree of agreement increases with deeper evaluation, the extrapolated value will go up; if agreement decreases, the extrapolated value will go down. Naturally, the extrapolated value is itself bounded by the upper and lower bounds of the RBO range. We have also proved that the distance measure $1 - \text{RBO}$ is a metric, and extended RBO in a consistent way to handle tied ranks and prefix rankings of different lengths. Finally, we have illustrated the use of RBO in comparing public search engines and in the IR researcher's laboratory, demonstrating that it gives stabler and more intuitive results than alternative measures.

Rank-biased overlap can properly be considered as a branch of a family of measures on indefinite rankings, which are overlap-based measures using a convergent set of weights over prefixes. We have argued that an overlap-based measure makes more sense for indefinite rankings than do measures derived from the notion of correlation. Indeed, our illustrative examples suggest that, in the presence of high and variable degrees of non-conjointness, correlation-based metrics tend in practice to degenerate into unweighted measures of set overlap.

While we have developed and deployed RBO in the IR field, it is applicable to any environment in which indefinite rankings occur – and these environments are numerous. Correct measurement is fundamental to informative observation and experimental manipulation, and when dealing with the volume of data produced by the modern information economy, measures that inform rather than confound are essential. We hope that rank-biased overlap will prove to be such a measure, for an important domain in which such measures have, until now, been lacking.

ACKNOWLEDGMENTS

We are grateful to the anonymous referees for their thoughtful and valuable comments. Charlie Clarke (University of Waterloo) was a participant in an early discussion.

A. TAIL DOMINATES PREFIX IN AO

In this Appendix, we prove that the tails of infinite rankings dominates the heads in the calculation of AO.

Consider the weight given to each rank by the AO measure on lists of depth n. Rank 1 is contained in each of the n subsets. In the first subset, it determines the entire overlap; in the second subset, it determines half the overlap; in the third, a third of the overlap; and so forth. Therefore the weight of rank 1 is:

$$W_{\mathrm{AO}}(1,n) = \frac{1}{n}\left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1} + \frac{1}{n}\right) = \frac{1}{n}\sum_{d=1}^{n}\frac{1}{d} = \frac{H_n}{n} \tag{33}$$

where $H_n \approx \gamma + \ln n + 1/(2n)$ is the nth Harmonic number, and $\gamma = 0.52771\ldots$ is the Euler-Mascheroni constant (see Knuth [1997, Section 1.2.7]). It follows that $W_{\mathrm{AO}}(2,n) = (H_n - H_1)/n$, that $W_{\mathrm{AO}}(3,n) = (H_n - H_2)/n$, and in general:

$$W_{\mathrm{AO}}(i,n) = \frac{H_n - H_{(i-1)}}{n}\ . \tag{34}$$

If only the prefix $k < n$ elements are available for each list, then the $\{1,\ldots,k\}$ heads of each list have contributed to the similarity measure, but the $\{k+1,\ldots,n\}$ tails have not. The cumulative weight of the head is:

$$\begin{aligned}
W_{\mathrm{AO}}^{head} &= \sum_{i=1}^{k} W_{\mathrm{AO}}(i,n) = \frac{1}{n}\sum_{i=1}^{k}\left(H_n - H_{(i-1)}\right) \approx \frac{1}{n}\ln\frac{n^k}{(k-1)!} && (35)\\
&= \frac{1}{n}\left[\ln n^k - \ln(k-1)!\right] \\
&\approx \frac{1}{n}\left[k\ln n - (k-1)\ln(k-1) + k - 1\right] && (36)\\
&\approx \frac{k}{n}\left[\ln n - \ln k + 1\right] \\
&= \frac{k}{n}\ln\frac{n}{k} + \frac{k}{n} && (37)
\end{aligned}$$

where the simplification at Equation 36 uses Stirling's approximation, $\ln x! \approx x\ln x - x$. Equation 37 goes to 0 as $n \to \infty$ and k is fixed.

The cumulative weight of the tail, following a similar line of simplification, is:

$$\begin{aligned}
W_{\mathrm{AO}}^{tail} &= \sum_{i=k+1}^{n} W_{\mathrm{AO}}(i,n) = \frac{1}{n}\sum_{i=k+1}^{n}\left(H_n - H_{(i-1)}\right) \approx \frac{1}{n}\ln\frac{n^{(n-k)}(k-1)!}{(n-1)!} && (38)\\
&\approx 1 - \frac{k}{n}\ln\frac{n}{k} - \frac{k}{n} && (39)
\end{aligned}$$

which goes to 1 as $n \to \infty$ and k is fixed. Therefore, for an infinite list, the weight of the tail is 1, and of the head is 0, proving that the tail dominates the head.

REFERENCES

BAR-ILAN, J. 2005. Comparing rankings of search results on the Web. *Information Processing & Management 41*, 1511–1519.

BAR-ILAN, J., MAT-HASSAN, M., AND LEVENE, M. 2006. Methods for comparing rankings of search engine results. *Computer Networks 50*, 10 (July), 1448–1463.

BLEST, D. C. 2000. Rank correlation – an alternative measure. *Australian and New Zealand Journal of Statistics 42*, 1, 101–111.

BUCKLEY, C. 2004. Topic prediction based on comparative retrieval rankings. In *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, Eds. Sheffield, United Kingdom, 506–507.

CARTERETTE, B. 2009. On rank correlation and the distance between rankings. In *Proc. 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, J. Allan, J. Aslam, M. Sanderson, C. Zhai, and J. Zobel, Eds. Boston, USA, 436–443.

CLIFF, N. 1996. *Ordinal methods for behavioural data analysis*. Lawrence Erlbaum Associates.

FAGIN, R., KUMAR, R., AND SIVAKUMAR, D. 2003. Comparing top k lists. *SIAM Journal on Discrete Mathematics 17*, 1, 134–160.

GIBBONS, J. D. AND CHAKRABORTI, S. 2003. *Nonparametric Statistical Inference*, 4th ed. CRC.

GOODMAN, L. A. AND KRUSKAL, W. H. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association 49*, 268, 732–764.

IMAN, R. L. AND CONOVER, W. J. 1987. A measure of top-down correlation. *Technometrics 29*, 351–357.

JÄRVELIN, K. AND KEKÄLÄINEN, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems 20*, 4, 422–446.

KENDALL, M. G. 1948. *Rank Correlation Methods*, 1st ed. Charles Griffin, London.

KNUTH, D. E. 1997. *The Art of Computer Programming, Vol. I: Fundamental Algorithms*, 3rd ed. Addison Wesley, Reading, MA. First edition 1968.

LESTER, N., MOFFAT, A., WEBBER, W., AND ZOBEL, J. 2005. Space-limited ranked query evaluation using adaptive pruning. In *Proc. 6th International Conference on Web Informations Systems Engineering*, A. H. Ngu, M. Kitsuregawa, E. J. Neuhold, J.-Y. Chung, and Q. Z. Sheng, Eds. New York, 470–477.

MELUCCI, M. 2007. On rank correlation in information retrieval evaluation. *SIGIR Forum 41*, 1, 18–33.

MELUCCI, M. 2009. Weighted rank correlation in information retrieval evaluation. In *Proc. 5th Asia Information Retrieval Symposium*, G. G. Lee, D. Song, C.-Y. Lin, A. Aizawa, K. Kuriyama, M. Yoshioka, and T. Sakai, Eds. Lecture Notes in Computer Science, vol. 5839. Sapporo, Japan, 75–86.

MOFFAT, A. AND ZOBEL, J. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems 27*, 1, 1–27.

QUADE, D. AND SALAMA, I. A. 1992. A survey of weighted rank correlation. In *Order Statistics and Nonparametrics: Theory and Applications*, P. K. Sen and I. A. Salama, Eds. Elsevier, 213–224.

SHIEH, G. S. 1998. A weighted Kendall's tau statistic. *Statistics & Probability Letters 39*, 17–24.

TARSITANO, A. 2002. Nonlinear rank correlation. Departmental working paper, Universitò degli studi della Calabria.

WU, S. AND CRESTANI, F. 2003. Methods for ranking information retrieval systems without relevance judgments. In *Proc. 18th Annual ACM Symposium on Applied Computing*. Melbourne, Florida, US, 811–816.

YILMAZ, E., ASLAM, J. A., AND ROBERTSON, S. 2008. A new rank correlation coefficient for information retrieval. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, Eds. Singapore, 587–594.

ZHAI, C. AND LAFFERTY, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems 22*, 2, 179–214.

Received March 2009; revised October 2009, January 2010; accepted March 2010.